The University of Jordan

Authorization Form

I, Manal Abdullah AbdelJabbar, authorize the University of Jordan to supply copies of my Thesis/ Dissertation to libraries or establishments or individuals on request, according to the University of Jordan regulations.

Signature:

Date: 22/ 5/ 2011

الجامعة الأردنيـــــة
كلية الدراسات العليا

التاريــــخ: / /

نموذج رقم (١٨)
اقرار والتزام بالمعايير الأخلاقية والأمانة العلمية
وقوانين الجامعة الأردنية وأنظمتها وتعليماتها
لطلبة الماجستير

أنا الطالب: منال عبدالله عبد عبد الجبار الرقم الجامعي: (8080983)

تخصص: أنظمة المعلومات الكليـــــة: الملك عبدالله الثاني لتكنولوجيا المعلومات

عنوان الرسالة: ....

Diagnosing hepatitis disease using data mining techniques

....

....

اعلن بأنني قد التزمت بقوانين الجامعة الأردنية وأنظمتها وتعليماتها وقراراتها السارية المفعول المتعلقة باعداد رسائل الماجستير عندما قمت شخصيا" باعداد رسالتي وذلك بما ينسجم مع الأمانة العلمية وكافة المعايير الأخلاقية المتعارف عليها في كتابة الرسائل العلمية. كما أنني أعلن بأن رسالتي هذه غير منقولة أو مستلة من رسائل أو كتب أو أبحاث أو أي منشورات علمية تم نشرها أو تخزينها في أي وسيلة اعلامية، وتأسيسا" على ما تقدم فانني أتحمل المسؤولية بأنواعها كافة فيما لو تبين غير ذلك بما فيه حق مجلس العمداء في الجامعة الأردنية بالغاء قرار منحي الدرجة العلمية التي حصلت عليها وسحب شهادة التخرج مني بعد صدورها دون أن يكون لي أي حق في التظلم أو الاعتراض أو الطعن بأي صورة كانت في القرار الصادر عن مجلس العمداء بهذا الصدد.

توقيع الطالب: منال .... التاريخ: ٢٢ / ٥ / ٢٠١١

تعتمد كلية الدراسات العليا
هذه النسخة من الرسالـــة
التوقيع...... التاريخ ..

DIAGNOSING HEPATITIS DISEASE USING DATA MINING TECHNIQUES

By

Manal Abdullah AbdelJabbar

Supervisor

Dr. Ammar Huneiti

Co-Supervisor

Dr. Azmi Mahafzah, Prof.

This Thesis was submitted in Partial Fulfillment of the Requirements for the Master's Degree of Information Systems

Faculty of Graduate Studies

The University of Jordan

May, 2011

تعتمد كلية الدراسات العليا
هذه النسخة من الرسالة
التوقيع ..... التاريخ .....

II

COMMITTEE DECISION

This Thesis/Dissertation (Diagnosing hepatitis disease using data mining techniques) was successfully Defended and Approved on 8/5/2011.

<u>**Examination Committee**</u>	<u>**Signature**</u>
Dr. Ammar Huneiti (Supervisor) Assist. Prof. of Adaptive Hypermedia	------------------------
Dr. Azmi Mahafza (Co- Supervisor) Prof. of Microbiology and Immunology	------------------------
Dr. Amjad Hudaib. (Member) Assoc. Prof. of Software Engineering	------------------------
Dr. Thair Hamtini (Member) Assist. Prof. of E-Learning	------------------------
Dr. Ghassan Kanaan (Member) Prof. of Artificial Intelligence (Al Petra University)	------------------------

تعتمد كلية الدراسات العليا
هذه النسخة من الرسالة
التوقيع...... التاريخ......

DIDECATION

This thesis is dedicates to my great parents for their love and encouragement.

ACKNOWLEDGEMENTS

I would like to thank Dr **Ammar Huneiti,** my supervisor , for his guidance, patience and support.

Pro. Azmi Mahafzah, my Co-Advisor**,** for his time and advice.

My parents, Thank you for letting me believe that we could be whoever we wanted to be, do whatever we wanted to do, there were no limitations except our drive, ambition and dreams. Thank you for everything.

My Sister Asmaa , thank you for support and encouragement.

My lovely usband, thank you for your understanding, love and Optimism.

I also would like to thank my sweet friends **Tasneem Abu-Kabeer** and **Nesreen Hamad** , thank you for always being with me.

Najat And Zeina thank you for your friendship and support.

Regards,
Manal Abdullah

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

HCC	Hepatocellular carcinoma
HBV	Hepatitis B virus
ANN	Artificial Neural Network.
CDSS	Clinical Decision Support System.
Rprop	Resilient Back Propagation Algorithm.
MLP	Multilayer Perceptron.
CI	Computational Intelligence.
ML	Machine Learning.
SOM	Self Organizing Maps.
TANH	Hyperbolic Tangent Sigmoid Function.
CDSSs	Clinical Decision Support Systems.
FS	Feature selection
T_BIL	Total bilirubin
TP	Total protein
CHE	Cholinesterase
GOT	Glutamic-oxaliacetic transminase
GPT	Glutmic-pyruvic transminase
LF	Linear Function.
RBF	Radial Basis Network.
PNN	Probabilistic Neural Network.
GRNN	General Regression Neural Network.

DIAGNOSING HEPATITIS DISEASE USING DATA MINING TECHNIQUES

By
Manal Abdullah AbdelJabbar

Supervisor
Dr. Ammar Huneiti

Co- Supervisor

Dr. Azmi Mahafzah, Prof

ABSTRACT

Extraction of meaningful information from large data sets is a key issue in the medical domain. In this study, a data mining framework is introduced to classify patients according to the type of viral hepatitis they have B and C. Hepatitis dataset which is provided by Chiba University Japan was studied. It contains laboratory tests of patients infected by hepatitis B and C.

To classify the type of virus, once the data have been prepared and encoded to a suitable format, two classification techniques were tested. The first technique is done using K means clustering technique which yields 84.70 % classification accuracy. The second is using multilayer perceptron (MLP) feed forward neural network .The input layer of the system includes five input variables; which are Total bilirubin, Total protein, Glutamic-pyruvic transminase, Glutamic-oxaliacetic transminase and Cholinesterase. The output layer contains one neuron that will produce a number that represents the hepatitis type. The number of hidden layers and nodes in the hidden layers are determined through a cascade learning process. Furthermore, a resilient back propagation (Rprop) algorithm is used to train the system. Experiments were performed with the proposed system. A classification accuracy of 85.29% was obtained from the experiments made on data taken from medical records of 340 patient suffering from chronic hepatitis.

1. INTRODUCION

1.1 Overview

Hepatitis as a disease is an inflammation of the liver that is commonly caused by a virus, particularly one of the five hepatitis viruses A, B, C, D, or E. This disease can be acute (short-lived) or chronic (lasting more than six months). Patients with chronic hepatitis treated for long time and large amount of information including medical tests is available. This medical information is considered an important source of knowledge.

Nowadays data mining techniques plays a great role in the medical domain. Extracting useful information and discovering the structure and relationships within these large amounts of data may help in diagnosing diseases. Disease diagnoses is a major challenge in the medical science. This has given rise to computerized medical tools, which are proposed to aid the physician in taking medical decisions. These computerized tools are called clinical decision support systems (CDSSs) which are computer programs that can be developed through several kind of machine learning (ML) techniques and used to assist clinicians care (Yan et al., 2006).

In this study it is attempted to discover differences in patterns between chronic hepatitis B and C based on laboratory test results. In other words patients are classified depending on their tests results. K means clustering and Artificial Neural network are the techniques that had been used to achieve these goals.

1.2 Problem Definition

A great responsibility lies on physicians to make a correct decision about what type of disease the patient has. Especially when dealing with chronic diseases like hepatitis that lasts more than six months and may persist for years. Huge amount of information stored for each patient including laboratory tests. These tests are done at different time intervals that vary from one day to several months. The major problems in this regard are determining the tests that affect the diagnosing and standardizing time interval for the tests.

The purpose of this study is to classify chronic hepatitis patients B and C based on medical tests results. In order to accomplish this, two classification techniques were used including K- means and Artificial Neural Network (ANN).

1.3 Proposed Techniques

The data set used in this study contains raw data in textual format that cannot be used as is. Preprocessing techniques were introduced to prepare the dataset. Firstly, features were selected with standard interval and the data were converted into suitable format. Two classification techniques are applied including is K means clustering which is an unsupervised learning algorithm. The second is Artificial Neural Network with a resilient back propagation. This thesis also investigates the effect of the number of hidden neurons and hidden layers in order to observe their impact on the classification process.

1.4 Organization of The Thesis

This thesis is organized as follows: chapter two gives literature review about thesis topic. Chapter three describes the research methodology and system architecture. The experimental results and analysis outlined in chapter four. Finally, the conclusion and future work are represented in chapter five.

2. LITRETURE REVIEW

2.1 Chapter Outline

This chapter gives background and some of the previous work related to thesis topic. It is organized as follows, an overview about data mining process is presented in section 2.2 In section 2.3,the machine learning process is discussed with its two types; supervised and unsupervised learning. Section 2.4 gives background about ANN. Information about hepatitis disease will be presented in section 2.5.

2.2 Data Mining

With huge amount of information stored in files, databases and other repositories it is becomes an important task to extract useful information from theses repositories. Data mining is the process of extraction information from large databases.

Data Mining is an analytical process, the main task of which is to explore large amount of data in order to discover consistent patterns and find relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. Data mining main tasks are prediction and description (Han and Kamber, 2000).

Prediction involves using some variables or fields in the data set to predict future value of other variables. On other hand, Description involves finding patterns describing the data that can be interpreted by humans (Fayyad, et al. 1996).

Data mining process involves five main steps (Han and Kamber, 2000); as shown in figure 1.

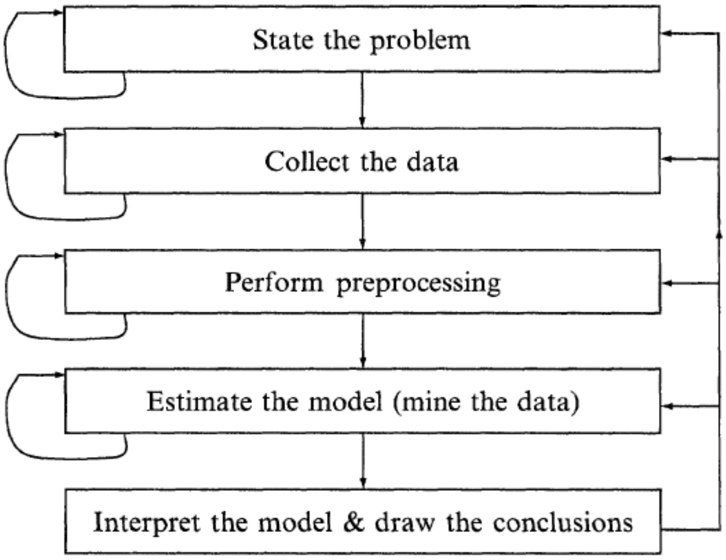


Figure 1: Data Mining Steps (Han and Kamber, 2000)

Larose. DT (2006) described the most common techniques used in data mining which are

1. Classification : Classify data set items into one of several classes
2. Regression : Maps a data item into a real value prediction variable
3. Clustering: Identify a finite set of categories or clusters to describe the data.
4. Summarization: Involves methods for finding a compact description for a set of data
5. Dependency Modeling: finding a local model that describes significant dependencies between variables or between the values of a feature in a data set or in a part of it.
6. Change and Deviation Detection: discovering the most significant changes in the data set.

2.3 Machine Learning

Is the area of artificial intelligence concerned with the study of computer algorithms that improve automatically through experience. Machine learning (ML) is the study of methods for programming computers to learn (Martine Swell, 2007).

Machine Learning techniques have been employed by many disciplines to automate complex decision making and problem solving tasks. This field of research is immensely diverse, with applications in medicine, mathematics, computer science, chemistry, economics, business management and many other fields (Ramalingam et al., 2006).

The majority of these techniques are concerned with pattern recognition problems. Pattern recognition is the act of taking in raw data and taking an action based on the "category" or the pattern (Duda et al., 2005), examples of pattern recognition problems are classification and regression (Mitchell, 1997). In other words; Machine learning can be defined as an attempt to automate doing science.

There are two types of machine learning (Martine Swell, 2007)

1. Supervised learning
2. Unsupervised learning

2.3.1 Supervised Learning

It is aimed to learn a mapping from the input to the output; correct values are available. In supervised learning, the user assembles a set of training data. The training data contains examples of inputs together with the corresponding outputs, and the network learns to infer the relationship between the two (Hagan , et al.,1996). Training data is usually taken from historical records. On other words, supervised learning is a learning

process in which changes in a network's weights and biases are due to the intervention of any external teacher (Freeman and Skapura ,1991).

Supervised learning has two steps

1. Learning (training): Learn a model using the training data

2. Testing: Test the model using unseen test data to assess the model accuracy

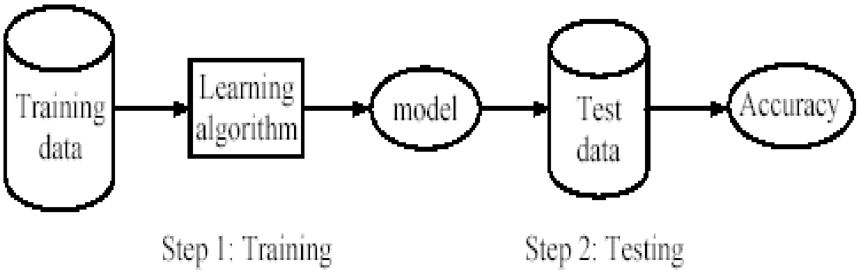


Figure 2: Supervised Learning Steps

A common supervised ML tools are; multilayer perceptron (MLP) neural network (Yan et al., 2006) and decision trees (Mitchell, 1997).

2.3.2 Unsupervised Learning

The main task is to establish the existence of classes or clusters in the data.

One form of unsupervised learning is clustering which is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Among neural network models, the Self-organizing map (SOM) is an example of unsupervised ML tool (Geoffrey and Terrence,1999).

2.3.2.1 Clustering

Clustering is the unsupervised classification of patterns into groups (clusters). Clustering is useful in several exploratory pattern-analysis, grouping, decision- making.

Typical pattern clustering activity involves the following steps (Jain and Dubes, 1988):

(1) Pattern representation (optionally including feature extraction and/or selection),

(2) Definition of a pattern proximity measure appropriate to the data domain,

(3) Clustering or grouping,

(4) Data abstraction (if needed), and

(5) Assessment of output (if needed).

Figure 3 shows clustering stages

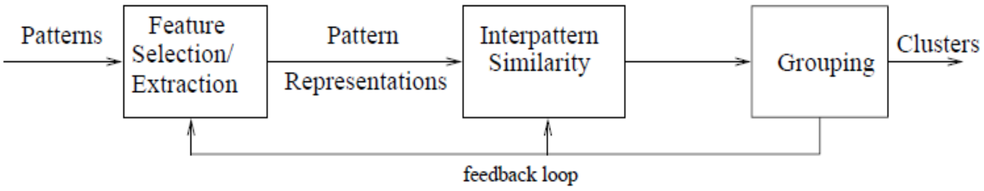


Figure 3: Clustering Stages (Jain and Dubes, 1988)

K-means is one of the simplest unsupervised learning algorithms that solve clustering problems. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point it is needed to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After having these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop it may be notice of that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (1)$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centers..

The algorithm is shown below (Jain and Dubes, 1988)

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Silhouette plot t can gives an idea of how well-separated the resulting clusters are. Silhouette value for each point is a measure of how similar that point is to points in its own cluster compared to points in other clusters, and ranges from -1 to +1. Silhouette value +1 indictes that the points are very distant from neighboring cluster, 0 indelicate

that points are not distinctly in one cluster or another , finally -2 indicate the points are probably assigned to the wrong cluster.

MacQueen, J. B. (1967) There are different types of distance functions that can be used in k-means for different datasets:

Euclidean : Each centroid is the function of the points in that cluster

Cityblock: Sum of absolute differences, i.e., L1. Each centroid is the component-wise median of the points in that cluster.

Consine : One minus the cosine of the included angle between points.

Each centroid is the mean of the points in that cluster, after normalizing those points to Euclidean length.

2.4 Artificial Neural Networks Background

Computational Intelligence **(CI)** is a set of nature-inspired computational methodologies and approaches to address complex problems of the real world applications to which traditional methodologies and approaches are ineffective or infeasible (Thammano and Moolwong, 2009). One of the techniques that belong to the CI family is Artificial Neural Networks **(ANNs)** (usually called Neural Network **(NN)**) that were introduced in 1940s (further developed in 1980s) as an attempt to emulate the features of biological neural networks in order to address a range of difficult information processing, analysis and modeling problems (Fausett, 2001). They are a powerful data modeling tool that is able to capture and represent complex input/output relationships. Since the motivation for the development of NN technology stemmed from the desire to develop an artificial system that could perform "intelligent" tasks

similar to those performed by the human brain; as a result neural networks resemble the human brain in the following two ways:

1. A neural network acquires knowledge through learning (experience).

2. A neural network's knowledge is stored within inter-neuron connection strengths known as weights.

ANNs have seen an explosion of interest over the last few years, and are being successfully applied across an extraordinary range of problem domains, in areas as diverse as finance, medicine, engineering, geology and physics (Mehrotra et al., 1996). Indeed, anywhere there are problems of prediction or classification, neural networks are being introduced (Haykin, 2001). This sweeping success can be attributed to a few key factors:

- **Power:** Neural network models can implicitly detect complex non linear relationships between independent and dependent variables, also they have the ability to detect all possible interactions between predictor variables and can be developed using multiple different training algorithms (Fausett, 2001).
- **Ease of use:** Neural networks learn by example. The neural network user gathers representative data, and then invokes training algorithms to automatically learn the structure of the data (Haykin, 2001). Although the user does need to have some heuristic knowledge of how to select and prepare data, how to select an appropriate neural network, and how to interpret the results.

Neural networks are applicable in virtually every situation in which a relationship between the predictor variables (independents, inputs) and predicted

variables (dependents, outputs) exists, even when that relationship is very complex and not easy to articulate in the usual terms of "correlations" or "differences between groups" (Fausett, 2001). A few representative examples of problems to which neural network analysis has been applied successfully are:

- **Credit assignment:** A variety of pieces of information are usually known about an applicant for a loan. For instance, the applicant's age, education, occupation, and many other facts may be available. After training a neural network on historical data, neural network analysis can identify the most relevant characteristics and use those to classify applicants as good or bad credit risks (Haykin, 2001).
- **Monitoring the condition of machinery:** Neural networks can be instrumental in cutting costs by bringing additional expertise to scheduling the preventive maintenance of machines (Mehrotra et al., 1996). A neural network can be trained to distinguish between the sounds a machine makes when it is running normally ("false alarms") versus when it is on the verge of a problem. After this training period, the expertise of the network can be used to warn a technician of an upcoming breakdown, before it occurs and causes costly unforeseen "downtime".

Due to the importance of ANNs as discussed above, the next subsections is concerned with giving a general background of the biological model that inspired the development of ANNs and comparing it with the basic model of an Artificial Neural Network **(ANN)**. Also the activation functions that are commonly used when building

an ANN model are also discussed. Beside the common two types of ANNs and the training algorithms, the designing issues that are related to ANNs are also presented.

2.4.1 The Biological Model

The basic model of the artificial neuron is founded upon the functionality of the biological neuron. By definition, “Neurons are basic signaling units of the nervous system of a living being in which each neuron is a discrete cell whose several processes are from its cell body (Freeman and Skapura, 1991).Each neuron is a specialized cell which can propagate an electrochemical signal. The neuron has a branching input structure (the dendrites), a cell body, and a branching output structure (the axon). The axons of one cell connect to the dendrites of another via a synapse as shown in Figure 4. When a neuron is activated, it fires an electrochemical signal along the axon. This signal crosses the synapses to other neurons, which may in turn fire. A neuron fires only if the total signal received at the cell body from the dendrites exceeds a certain level (the firing threshold). The strength of the signal received by a neuron (and therefore its chances of firing) critically depends on the efficacy of the synapses. Each synapse actually contains a gap, with neurotransmitter chemicals poised to transmit a signal across the gap junction (Hagan et al., 1996).

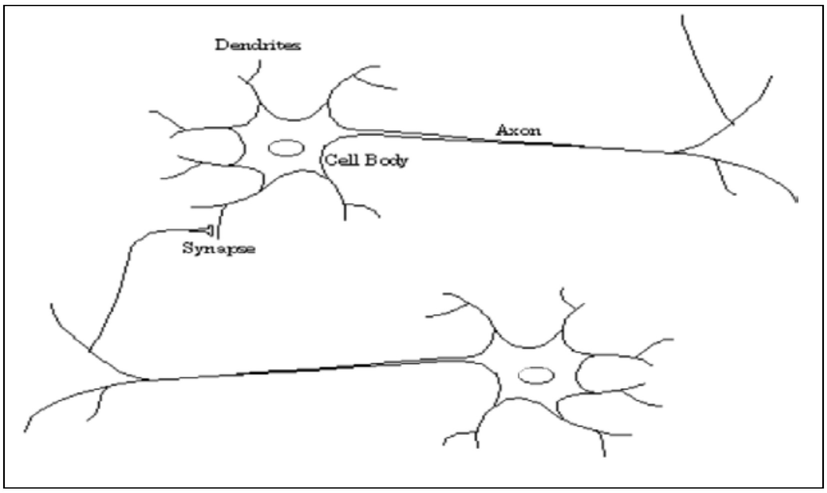


Figure 4: Biological Neuron; (Haykin, 2001)

The brain which is principally composed of a very large number (10,000,000,000) of neurons, massively interconnected (with an average of several thousand interconnects per neuron, although this varies enormously) will analyze all patterns of signals sent, and from that information it interprets the type of information received. Thus, from a very large number of neurons (each performing a weighted sum of its inputs, and then firing a binary signal if the total input exceeds a certain level) the brain manages to perform extremely complex tasks (Krose and Smagt, 1996). Of course, there is a great deal of complexity in the brain which has not been discussed here, but it is interesting that ANNs can achieve some remarkable results using a model not much more complex than this.

2.4.2 The Basic Model

The basic model of ANNs is an input layer consisting of nodes that simply accept the input values. The outputs of neurons in a layer are inputs to neurons in the next layer. The last layer is called the output layer. Layers between the input and output layers are known as hidden layers. Figure 5 illustrates a diagram for this architecture.

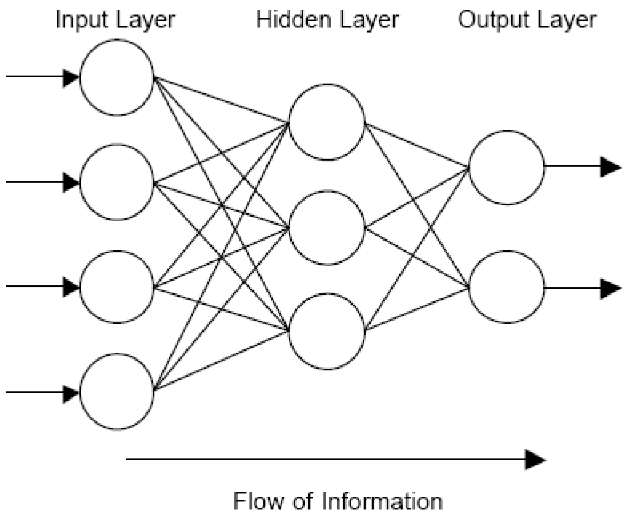


Figure 5: The basic model of an ANN

To capture the essence of biological neural systems, an artificial neuron is defined (Fausett, 2001) as follows:

- It receives a number of inputs (either from original data, or from the output of other neurons in the neural network). Each input comes via a connection that has a strength (or weight); these weights correspond to synaptic efficacy in a biological neuron.
- The weighted w_i sum of the inputs u_i is formed to compose the activation of the neuron. Gathering the weighs of the inputs according to:

$$Sum = \sum_{i=1}^{n} u_i w_i \quad (2)$$

- The activation signal is passed through an activation function *(y)* (also known as a transfer function) to produce the output of the neuron as follows:

$$y = f\left(\sum_{i=1}^{n} u_i w_i\right) \qquad (3)$$

Figures 6 illustrates a summary of the principle work of an artificial neuron model, and a schematic illustrations of the basic features of a biological neuron and its artificial basic model is illustrated in Figure 7

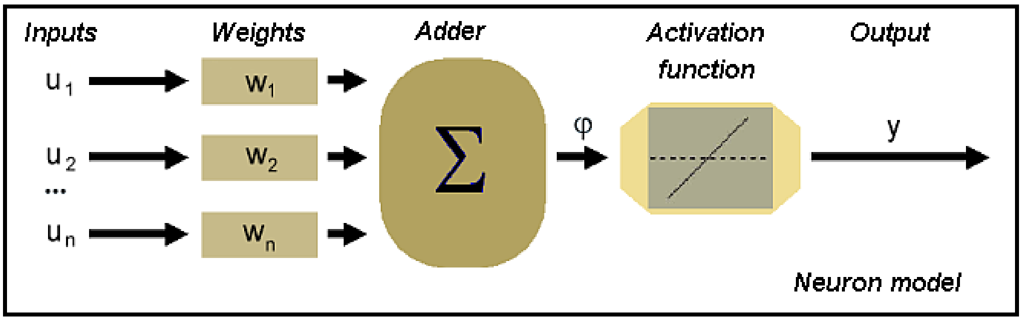


Figure 6:Summary of the principle work of Artificial Neuron Model; (Fausett, 2001)

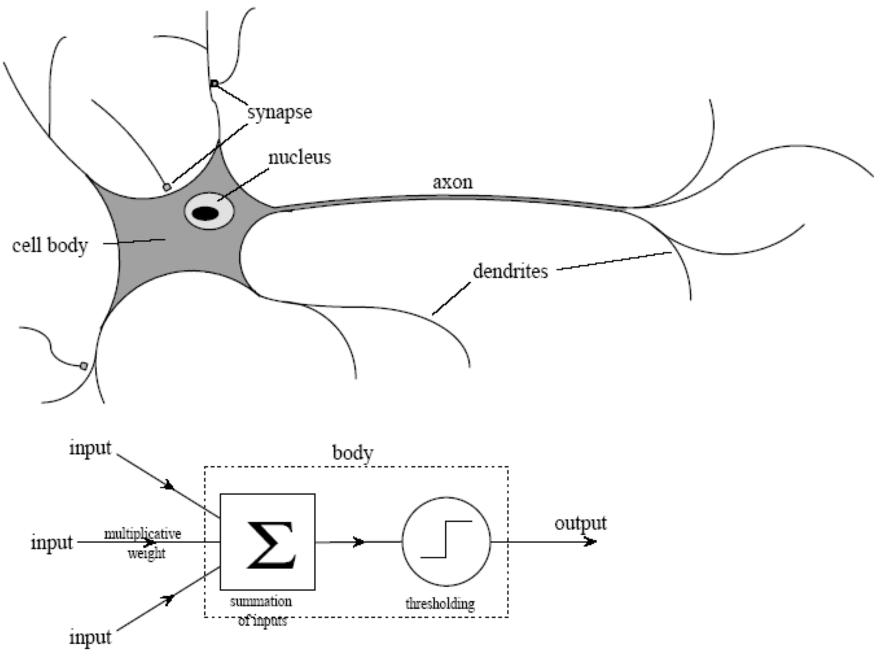


Figure 7: Schematic Illustrations of the Basic Features of a biological Neuron and its Artificial Basic Model; (Fausett, 2001)

2.4.3 Types of Activation Function

The output of a particular artificial neuron is determined by the type of activation function, known sometimes as transfer function (Freeman and Skapura, 1991).An activation function is used to limit the amplitude of the output of a neuron and is commonly known as a squashing function (Haykin, 2001). Neural Networks supports a wide range of activation functions. The most common types of activation functions are:

1. **Linear activation function (LF)**: The outputs of a linear activation function, also known as a linear combiner, are determined as follows: ***a* = *n*** where ***n*** is the sum of the weighted inputs. The graph of a LF function is represented as in Figure 8

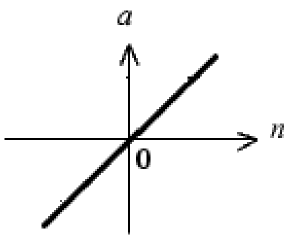


Figure 8: Graph of a Linear Activation Function

2. **Hyperbolic tangent sigmoid activation function (tanh)**: The hyperbolic tangent

sigmoid activation function (tanh) has the following formula: $a = \frac{2}{1+e^{-2n}} - 1$

where *n* is the sum of the weighted inputs. The resulted output from this function will be in a range between [-1,1] (Hagan et al., 1996). The graph of a tanh function is represented as in Figure 9

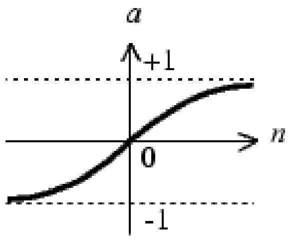


Figure 9: Hyperbolic Tangent Sigmoid Activation Function

2.4.4 ANNs Architectures

ANNs architecture can be divided into two main categories based on their processing structure (known as topology): feedback and feed-forward neural networks.

In feedback neural networks, signals can travel in both directions. Thus the network has loops in its topology as shown in Figure 10(a). Networks with this structure can be very complicated and powerful, as they can continuously change state until an equilibrium is reached – this balanced state is then kept up until changes in the input data occur (Mehrotra et al., 1996).

In feed-forward neural networks signals are allowed to travel one way only; from input to output as shown in Figure 10 (b). There are no feedback (loop) connections. Feed-forward networks tend to be straight forward models that associate inputs with outputs (Mehrotra et al., 1996).

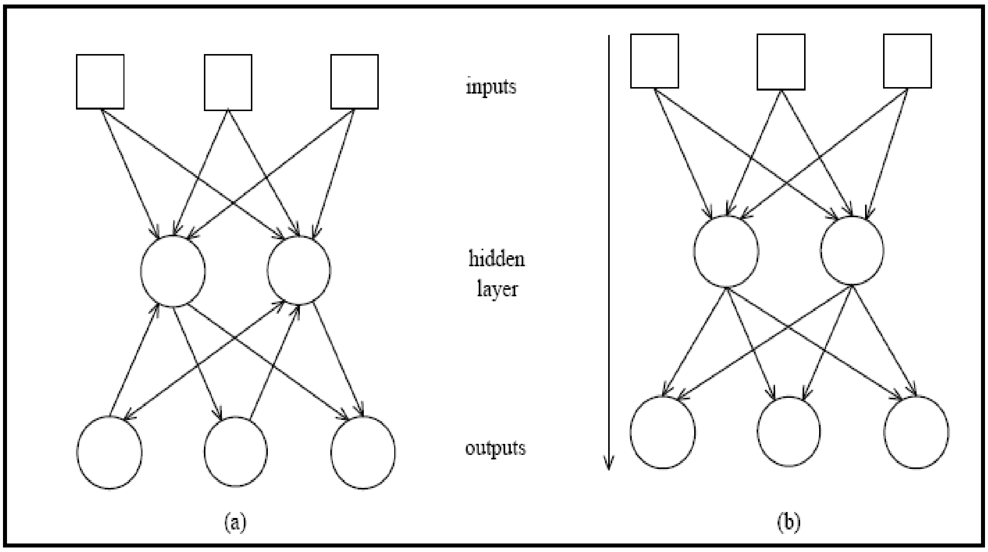


Figure 10: A (a) Feedback and a (b) Feed-forward Neural Network; (Mehrotra et al., 1996).

These architectures can be further subdivided into two main classes based on the number of layers of processing nodes used in the model; single-layer perceptron and multi-layer perceptron (Hagan et al., 1996). In a single-layer perceptron there is a layer of input nodes and a layer of output nodes, which are the only processing nodes in the model. In a multi-layer perceptron (MLP) there is one or more hidden layers between the input and the output nodes.

2.4.5 The Back Propagation Algorithm

MLP neural networks have been applied successfully to solve difficult and diverse problems by training them in a supervised manner with highly popular algorithm known as Back Propagation which uses the data to adjust the network's weights and biases in a manner that minimizes the error in its predictions on the training set (Krose and Smagt, 1996). If the network is properly trained, it has then learned to model the (unknown) function that relates the input variables to the output variables, and can subsequently be used to make predictions where the output is not known.

A Back Propagation network learns by example. The algorithm is given examples of what the network is needed to do and it changes the network's weights so that, when training is finished, it will give you the required output for a particular input. In order to train the network it is needed to be given it examples of what output needed (called the *Target*)) for a particular input as shown in Figure 11.

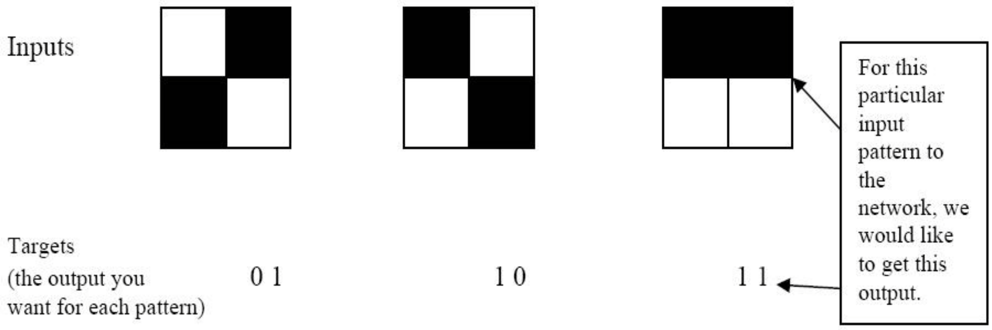


Figure 11: A Back Propagation Training Set; (Krose and Smagt, 1996)

So, the first pattern is put into the network, we would like the output is needed to be 0 1 as shown in Figure 12 (a black pixel is represented by 1 and a white by 0). The input and its corresponding target are called a *Training Pair*. Once the network is trained, it will provide the desired output for any of the input patterns.

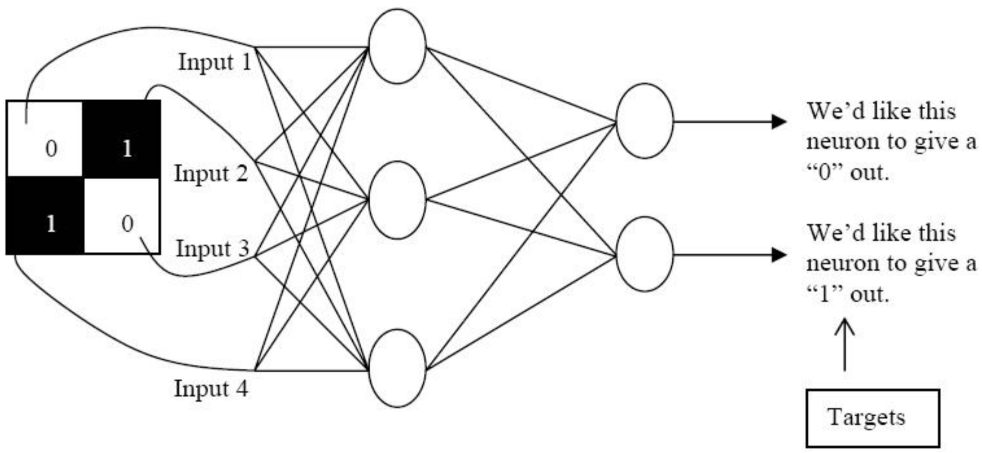


Figure 12: Applying a Training Pair to a Network; (Krose and Smagt, 1996)

The training works as follows (Krose and Smagt, 1996):

- The network is first initialised by setting up all its weights to be small random numbers.
- Next, the input pattern is applied and the output calculated (this is called the *forward pass*). The calculation gives an output which is completely different to what you want (the Target), since all the weights are random.
- The *Error* of each neuron, is then calculated which is essentially:

  Target – Actual Output (i.e. What you want – What you actually get).
- This error is then used mathematically to change the weights in such a way that the error will get smaller. In other words, the Output of each neuron will get closer to its Target (this part is called the *reverse pass*).
- The process is repeated again and again until the error is minimal.

2.4.5.1 The Error Surface

Since Back Propagation iteratively searches for a set of weights that minimizes the error function **(E)** over all training pairs, MLP networks have an error surface, which is in general complex and believed to have many local and global minima. This occurs because the algorithm always changes the weights in such a way as to cause the error to fall. But the error might briefly have to rise, as shown in Figure 13. If this is the case, the algorithm will get stuck (because it can’t go uphill) and the error will not decrease further (Fausett, 2001). This requires rerunning the algorithm with new random initial weights and other network parameter hoping to reach the global minima in the new run. As a consequence, Back Propagation is never assured to find the global minima.

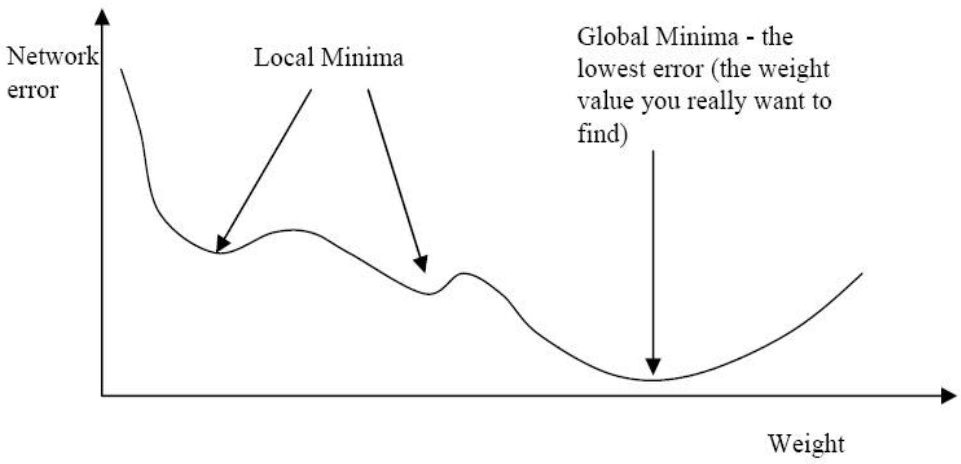


Figure 13: An Error Surface

2.4.5.2 The learning Rate

An important consideration in the Back Propagation algorithm is the learning rate ***η*,** which determines by how much we change the weights ***w*** change at each step. If ***η*** is too small, the algorithm will take a long time to converge (reaching an optimal set of weights that give the solution) (Hagan et al.,1996). The Influence of choosing a small learning rate on the error surface is illustrated in Figure 14

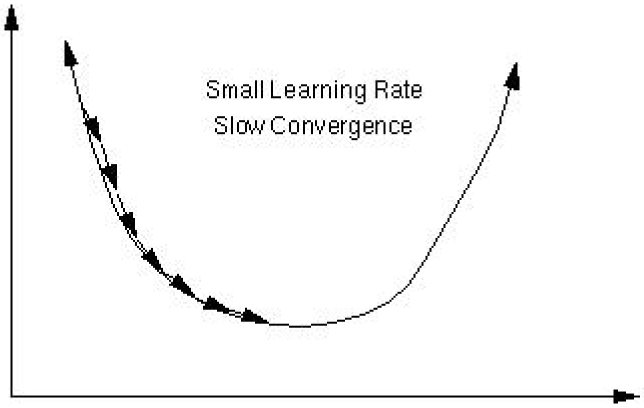


Figure 14: The Influence of Choosing a Small Learning Rate on the Error Surface

Conversely, if η is too large, we may end up bouncing around the error surface out of control, which make the algorithm to diverge from the solution. The influence of choosing a large learning rate is illustrated in Figure 15.

The correct setting for the learning rate is application-dependent, and is typically chosen by experiment; it may also be time-varying, getting smaller as the algorithm progresses (Hagan et al.,1996).

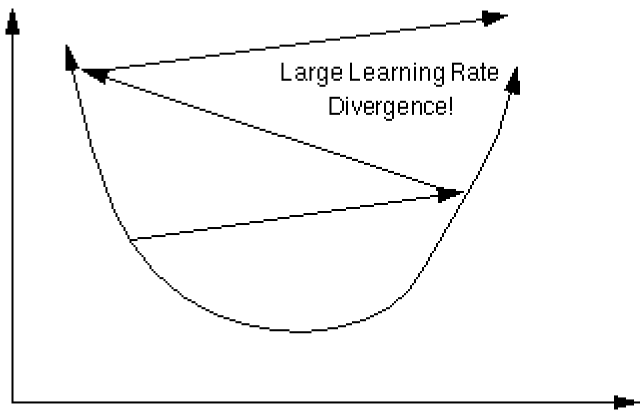


Figure 15: The Influence of Choosing a Large Learning Rate on the Error Surface

2.4.5.3 The Back Propagation Steps in Details

The architecture of Backpropagation as shown in Figure 16 work accordingly to the following: (Freeman and Skapura ,1991).

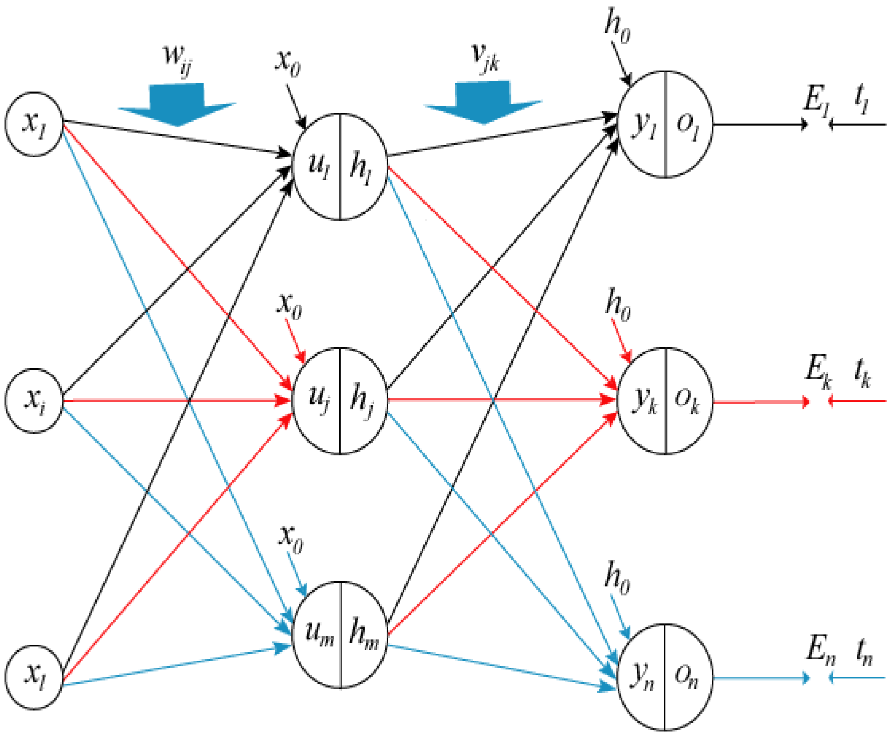


Figure 16: Architecture of Back Propagation

Step 1: Initialize the weights of the randomly with numbers between -1 and 1.

$$w_{ij} = random([-1,+1]) \qquad 0 \le i \le l, 1 \le j \le m \tag{4}$$

$$v_{ij} = random([-1,+1]) \qquad 0 \le j \le m, 1 \le k \le n \tag{5}$$

Step 2: Present the pattern *p* to the network, where p is one pattern from the whole pattern P:

$$p = \left(x_1, x_2, x_3, x_i, ...x_l\right) \in R^l \qquad p \in P \qquad (5)$$

Step 3: Compute the values of the hidden-layer nodes using the formula:

$$u_j = \sum_{i=0}^{l} w_{ij} \times x_i \qquad , h_j = f(u_j) \qquad ,1 \le j \le m \qquad (6)$$

Step 4: Calculate the values of the output nodes using the formula:

$$y_k = \sum_{j=0}^{m} v_{ik} \times h_j \qquad , o_k = f(y_k) \qquad ,1 \le k \le n \qquad (7)$$

Step 5: Compute the errors in the output layer using:

$$\delta o_k = (t_k - o_k) o_k (1 - o_k) \qquad ,1 \le k \le n \qquad (8)$$

Step 6: Compute the errors in the hidden layer using:

$$\delta h_j = h_j (1 - h_j) \sum_{k=1}^{n} \delta o_k \times v_{jk} \qquad ,1 \le j \le m \qquad (9)$$

Step 7: Adjust the weights between the output layer and the hidden according to:

$$v_{jk}^{(new)} = v_{jk}^{(old)} + \eta \times \delta o_k \times h_j \qquad ,0 \le j \le m, 1 \le k \le \qquad (10)$$

Step 8: Adjust the weights between the hidden layer and the input layer according to

$$w_{ij}^{(new)} = w_{ij}^{(old)} + \eta \times \delta h_j \times x_i \qquad ,0 \le i \le l \qquad ,1 \le j \le m \quad (11)$$

Step 9: Repeat step 2 through 8 for each element of training set, until the Total Mean Squared Error **(TMSE)** ≤ E or reach the maximum , E is an error threshold that is used to stop the training, error E defined as the difference between the desired t_k and current o_k output where (Hagan et al.,1996) :

$$TMSE = \frac{1}{N} \sum_{k=1}^{n} (t_k - o_k)^2 \qquad (12)$$

We notice from step 7 and 8 that these steps are the ones that change the weights of the network based on the following general formula:

$$w_{ij}^{(new)} = w_{ij}^{(old)} + \Delta w_{ij} \qquad (13)$$

The term Δw_{ij} , is the amount of change in the current weight, this term in the back propagation algorithm is equal to; $\eta \times \delta o_i \times h_j$ where the term $\delta o_i \times h_j$ represent the partial derivative of the error with respect to the weight ($\frac{\partial E}{\partial w_{ij}}$) (Hagan et al., 1996).

The whole idea behind the partial derivative is; to gradually, but consistently, decrease the output error by adjusting the weights. The trick is to figure out how to adjust the weights. Intuitively, we know that if a change in a weight will increase the error, then we want to decrease that weight. Mathematically, this means that we look at

the derivative of the error with respect to the weight: $\frac{\partial E}{\partial w_{ij}}$, once we find this derivative, we will update the weight via the following:

$$\Delta w_{ij} = \eta \times \frac{\partial E}{\partial w_{ij}} \tag{14}$$

2.4.6 Resilient Back Propagation Algorithm (Rprop)

As discussed in subsection 2.1.5.1, a Back Propagation algorithm will experience a local minimum that occurs because the algorithm always changes the weights based on the value of the partial derivative. So in order to solve this problem; the Resilient Back Propagation **(Rprop)** was devised (Riedmiller and Braun, 1993), where it eliminates the harmful influence of the magnitude of the partial derivative. As a consequence, only the sign of the derivative is considered to indicate the amount of the weight update. To achieve this, an update-value Δ_{ij} is introduced for each weight $\boldsymbol{w_{ij}}$, which solely determines the size of the weight-update (Riedmiller and Braun, 1993). This adaptive update-value evolves during the learning process**,** according to the following learning-rule:

$$\Delta_{ij}^{(t)} = \begin{cases} \eta^{+} \times \Delta_{ij}^{(t-1)} & , \quad if \ \frac{\partial E}{\partial w_{ij}}^{(t-1)} \times \frac{\partial E}{\partial w_{ij}}^{(t)} \rangle \ 0 \\ \eta^{-} \times \Delta_{ij}^{(t-1)} & , \quad if \ \frac{\partial E}{\partial w_{ij}}^{(t-1)} \times \frac{\partial E}{\partial w_{ij}}^{(t)} \langle \ 0 \\ \Delta_{ij}^{(t-1)} & , \quad else \end{cases} \tag{15}$$

$$where \quad 0 \ \langle \ \eta^{-} \ \langle \ 1 \ \langle \ \eta^{+}$$

The adaptation-rule works as follows: Every time the partial derivative of the corresponding weight $\boldsymbol{w_{ij}}$ changes its sign, which indicates that the last update was too big and the algorithm has jumped over a local minimum, the update-value Δ_{ij} is decreased by the factor η^{-} (which was found by experiments that the best value for it is equal to **0.5**)**.** If the derivative retains its sign, the update-value is slightly increased by the factor η^{+} (which was found by experiments that the best value for it is equal to **1.2**) in order to accelerate convergence in shallow regions.

Once the update-value for each weight is adapted, the weight-update itself follows a very simple rule: if the derivative is positive, the weight is decreased by its update-value, if the derivative is negative, the update –value is added Riedmiller and Braun, 1993):

$$\Delta w_{ij}^{(t)} = \begin{cases} -\Delta_{ij}^{(t)}, & if\ \dfrac{\partial E}{\partial w_{ij}}^{(t)} \rangle\ 0 \\ +\Delta_{ij}^{(t)}, & if\ \dfrac{\partial E}{\partial w_{ij}}^{(t)} \langle\ 0 \\ 0, & else \end{cases} \tag{16}$$

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} + \Delta w_{ij}^{(t)} \tag{17}$$

In the Rprop the weights and update-values are changed every time the whole pattern set has been presented once to the network (learning by epoch). Therefore the term $\frac{\partial E}{\partial wij}^{(t)}$ denotes the summed gradient information over all patterns.

2.4.6.1 Rprop Algorithm

The following pseudo-code fragment shows the kernel of the **RPROP** adaptation and learning process. The **minimum (maximum)** operator is supposed to deliver minimum (maximum) of two numbers; the **sign** operator returns +1, if the argument is positive, -1, if the argument is negative, and 0 otherwise. (Freeman and Skapura ,1991).

For all weigths and biases {

if ($\frac{\partial E}{\partial w_{ij}}^{(t-1)} \times \frac{\partial E}{\partial w_{ij}}^{(t)} \rangle \ 0)$ **then** {

$$\Delta_{ij}^{(t)} = \textbf{minimum}\,(\Delta_{ij}^{(t-1)} \times \eta^{+} \quad , \Delta_{\max})$$

$$\Delta w_{ij}^{(t)} = \textbf{sign}\,(\frac{\partial E}{\partial w_{ij}}^{(t)}) \times \Delta_{ij}^{(t)}$$

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} + \Delta w_{ij}^{(t)}$$

}

else **if** ($\frac{\partial E}{\partial w_{ij}}^{(t-1)} \times \frac{\partial E}{\partial w_{ij}}^{(t)} \langle \ 0)$ **then** {

$$\Delta_{ij}^{(t)} = \textbf{maximum}\,(\Delta_{ij}^{(t-1)} \times \eta^{-} \quad , \Delta_{\min})$$

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} - \Delta w_{ij}^{(t-1)}$$

$$\frac{\partial E}{\partial w_{ij}}^{(t)} = 0$$

}

else **if** ($\frac{\partial E}{\partial w_{ij}}^{(t-1)} \times \frac{\partial E}{\partial w_{ij}}^{(t)} = \ 0)$ **then** {

$$\Delta w_{ij}^{(t)} = \textbf{sign}\,(\frac{\partial E}{\partial w_{ij}}^{(t)}) \times \Delta_{ij}^{(t)}$$

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} + \Delta w_{ij}^{(t)}$$

}

}

2.5 Diagnosing Hepatitis Disease

2.5.1 Hepatitis Disease Overview

Hepatitis means inflammation of the liver, liver is a vital organ that processes nutrients, filters the blood, and fights infections. When the liver is inflamed or damaged, its function can be affected. (Dienstag et al., 2001)

The inflammation can be acute, flaring up and then resolving within a few weeks to months, or chronic, enduring over many years. Chronic hepatitis may persist for 20 years or more before causing significant symptoms related to progressive liver damage such as cirrhosis, liver cancer, or death. There are many forms of hepatitis, including Hepatitis A, B and C. It is transmitted through infected blood and other body fluids and unsafe sexual practices. It can also be transmitted from mothers to babies. Chronic hepatitis B (CHB) affects 350 million individuals worldwide. Prenatal transmission leads to high rates of chronic infection and complications, including cirrhosis and hepatocellular carcinoma. (Peteres , 2007)

The signs and symptoms of hepatitis are the same, regardless of the cause, but vary from person to person and may vary over time. Many people have few, mild, and/or vague symptoms that may be mistaken for the flu (Locarnini,2004).Some of the more common signs and symptoms include; Fatigue, Nausea, Abdominal pain, Joint aches, Itching and Jaundice

Hepatocellular carcinoma (HCC) is the most common type of liver cancer and the fifth most common cancer in the world. The exact cause of HCC is unknown. Viruses such as hepatitis B and hepatitis C have been shown to increase the risk of HCC. (HO et al .2005)

2.5.2 Managing Hepatitis Disease

Shoji Hirano and Shusaku Tsumoto(2005) worked much on hepatitis data set .In (2005) they presented a method called "the modified multi scale matching". They applied the proposed method to chronic hepatitis patients. The method enables them to compare sequences by partly changing observation scales. In 2006 they continue their

work on chronic hepatitis dataset and proposed another cluster analysis method that deals with time-series medical data base on the trajectory.

This method represents the time series of test results as trajectories in multidimensional space, and compares their structural similarity by using the multiscale comparison technique. As a result, it is possible to get groups of cases that have correspondence to the fibrosis stages. In 2007 they presented a new approach to find knowledge from temporal data on chronic diseases based on the combination of advanced sequence comparison techniques and cluster analysis procedure.The result of temporal analysis aiming at finding years for reaching F4 (fibrosis stage 4) , years between stages, and their relationships with virus types and fibrosis stages. In 2008 they proposed a novel cluster analysis method for multivariate time-series medical data laboratory tests. Clustering experiments on chronic hepatic dataset shows that the data can be split into two categories ;one contains the progressed cases and the other contains un-progressed cases.

Nguyen et al. (2003) presented a temporal abstraction approach to mining the temporal hepatitis data. The basic idea of temporal abstraction methods to deal with long and irregular time-stamp sequences are the separation of long-term and short-term changed tests groups, and doing abstraction of each group in efficient and appropriate ways. This technique provides many advantages in mining temporal data.

Kenji et al (2007) presented a method for clustering time-series medical data based on Formal Concept Analysis. They used Formal Concept Analysis as a method to find similarities in time series data of medical tests. It is a useful approach to find meaningful patient clusters as well as classifying medical cases.

Rouhani and haghighi (2009) used vector machines and artificial neural network to diagnose hepatitis disease. They use this techniques to identify the type and the phase of the disease. The overall accuracy for their diagnostic system is 97%.

3. Methodology and System Architecture

3.1 Overview

In this study, a data mining framework, which includes choosing data set ,feature selection, data cleansing, data normalization, clustering and classification, is introduced. Figure 17 shows the overall research methodology that was followed throughout the study.

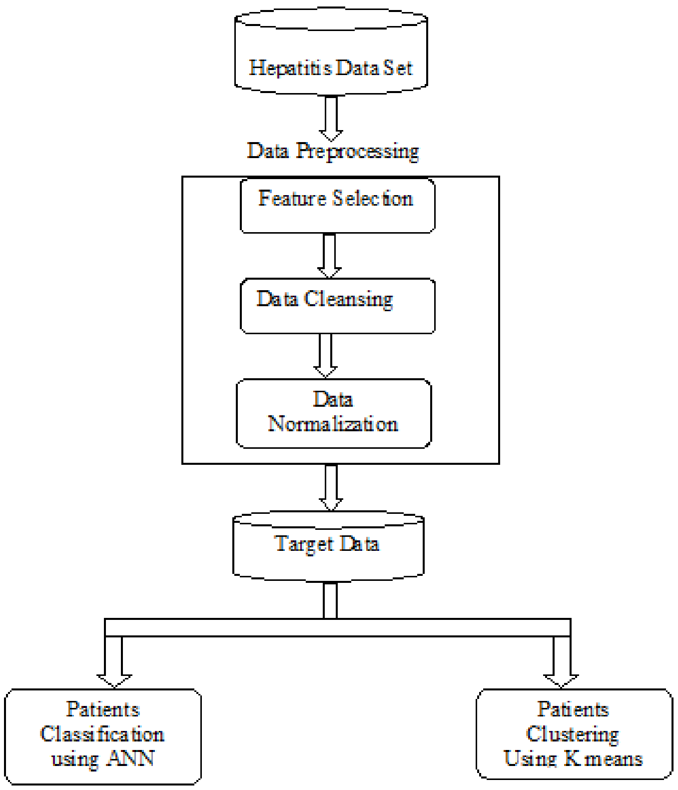


Figure 17: Research Methodology

The following subsections will explain all steps in details.

3.2 Data Preprocessing

Data preprocessing is an important phase in the data mining process and has a huge impact on the success of the data mining process. The real datasets usually contain noise and inconsistent data.

The data from the Hepatitis dataset is in CSV format; numbers and text are stored in textual form and can't be used directly for analysis without preprocessing. This dataset must be cleaned and transformed into suitable format in order to be introduced to the data classification techniques. In this section the steps that were followed to prepare the data set will be discussed. These are

- Feature selection
- Data cleansing
- Data normalization

After completing the preprocessing phase it is expected to have data with the following characteristics

- Data with smaller size
- Data in suitable format
- Clean data; i.e. eliminate noise and irrelevant records
- Data in uniform scale; the values of the tests is in the same range

3.2.1 Feature Selection

3.2.1.1 Overview

Feature selection is an essential part in the data mining process because the data set may contain many irrelevant variables that could be omitted. The main idea of feature

selection is to choose a subset of input variables by eliminating features with little or no predictive information (Cheng et. al , 2006).

It improves the performance of the learning algorithms, reduces the computational cost, and provides better understandings of the datasets.

An important challenging issue is the selection of appropriate subset of the available features so that the selected subset can adequately model the target task. In medical domain classification analysis applications generally take one of the following two feature selection approaches, including use of automatic feature selection mechanisms or expert judgment.

Automatic feature selection mechanisms: often use a search algorithm (such as hill climbing and best first) along with a heuristic for evaluating the goodness of a feature subset to search in the space of possible hypotheses and to identify important features in reasonable time Yoon et al. (1995).

The expert judgment approach feature selection: does not rely on the training dataset for selection decision. After several years of training and practices, experts in a subject domain generally develop some rules of thumb in problem solving and accumulate insights on new, rare, complex, or otherwise poorly understood phenomena.

Hepatitis data set contains approximately hundreds of laboratory tests, so we have to select the tests that detect the type hepatitis virus infection. Selecting the suitable features yield a better prediction and classification. On other hand including irrelevant features may produce less accurate and less understandable results (Lei and Huan , 2004).

3.2.1.2 Selecting Relevant Features:

Cheng et. al (2006) proved that the certain tests can be used to detect the hepatitis type.

These tests are **(i)** Total bilirubin (T_BIL), **(ii)** total protein (TP), **(iii)** cholinesterase (CHE), **(iv)** glutamic-pyruvic transminase (GPT) and **(v)** glutamic-oxaloacetic transminase(GOT) .

The selected tests are liver function test that are used to detect liver damage or disease. (Thapa and Walia , 2007)

- **Total bilirubin (T_BIL):**T_BIL is produced when the liver breaks down old red blood cells. Bilirubin is then removed from the body through the stool (feces) and gives stool its normal brown color.
  - **Indirect (or unconjugated) bilirubin.** This form of bilirubin does not dissolve in water (it is insoluble). Indirect bilirubin travels through the bloodstream to the liver, where it is changed into a soluble form (direct or conjugated).
  - **Direct (or conjugated) bilirubin.** Direct bilirubin dissolves in water (it is soluble) and is made by the liver from indirect bilirubin.

    Total bilirubin and direct bilirubin levels are measured directly in the blood, whereas indirect bilirubin levels are derived from the total and direct bilirubin measurements.

    When bilirubin levels are high, the skin and whites of the eyes may appear yellow (jaundice). Jaundice may be caused by liver disease (hepatitis),

- **Total Protein (TP):**

Protein is needed to build and maintain muscles, bones, blood, and organ tissue. Sometimes when there is a problem with the liver, it can't make proteins, so protein

levels decrease. This test measures the albumin and all other proteins in blood including antibodies made to help in fighting infections.

- **Glutamic Pyruvate Transaminase (GPT):**

Serum Glutamic Pyruvate Transaminase (SGPT), also called Alanine transaminase (ALT), is an enzyme present in hepatocytes (liver cells). When a cell is damaged, it leaks this enzyme into the blood, where it is measured. ALT rises dramatically in acute liver damage, such as viral hepatitis or paracetamol (acetaminophen) overdose. Elevations are often measured in multiples of the upper limit of normal (ULN).

because ALT is found primarily in the liver, high levels of ALT almost always indicate that there is a problem with the liver. (Conversely, a normal ALT level does not necessarily mean that the liver is definitely normal

- **Glutamic-oxaloacetic transminase(GOT):**

Serum Glutamic Oxaloacetic Transaminase (SGOT), also called Aspartate transaminase (AST), is raised in acute liver damage, but is also present in red blood cells, and cardiac and skeletal muscle and is therefore not specific to the liver. The ratio of AST to ALT is sometimes useful in differentiating between causes of liver damage.Elevated AST levels are not specific for liver damage, and AST has also been used as a cardiac marker.

Because AST is found in many other organs besides the liver, including the kidneys, the muscles, and the heart, having a high level of AST does not always (but often does) indicate that there is a liver problem.

- **Cholinesterase:**

Serum cholinesterase is a test that looks at blood levels of certain enzymes (acetylcholinesterase and pseudocholinesterase) that help the nervous system work properly.

Acetylcholinesterase is found in nerve tissue and red blood cells. Pseudocholinesterase is found primarily in the liver. This test may done infrequently to diagnose liver disease. This test was chosen because as mentioned above liver is the source of plasma cholinesterase and low plasma cholinesterase activity has been reported in patients with liver disease.

3.3 Data Cleansing

Data Cleansing is the process of ensuring that all values in a dataset are consistent and correctly recorded. This step is very important especially when dealing with datasets that contain large amount of laboratory test results. The data cleansing process is done using the following:

1. Choosing the standard interval for the selected features
2. Building patient/features matrix

3.3.1 Choosing Standard Interval for The Selected Attributes

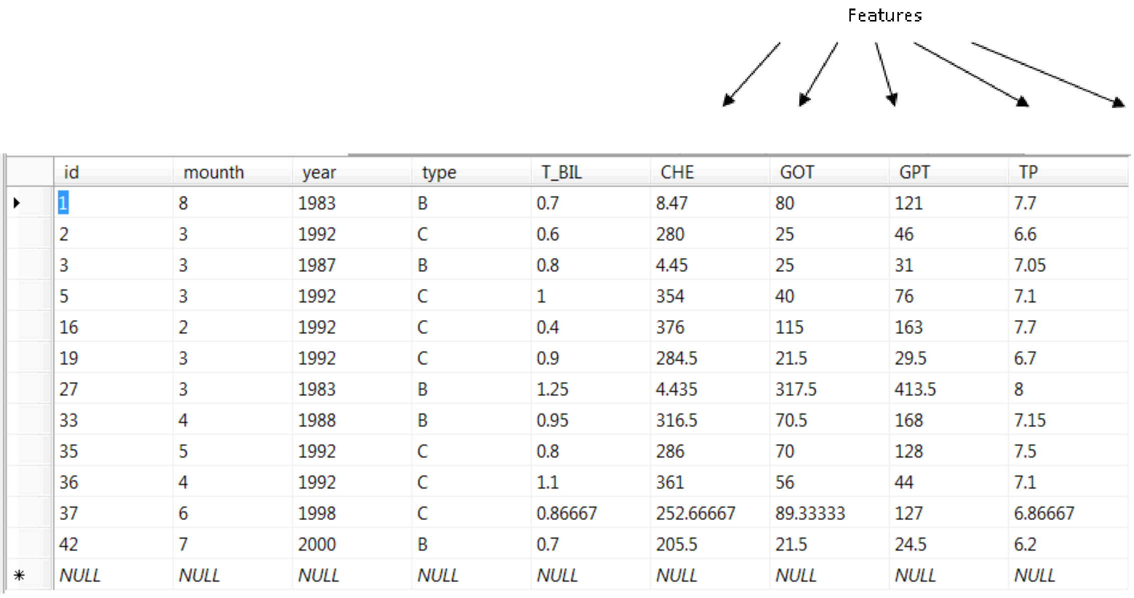
Patients are not tested regularly; period between two tests can range from one day to several months. In order to standardize information about each patient data that were collected during a specific well defined time interval of fixed length (ΔT) for each patient were selected. The considered interval begins in specific time. The specific time in this work is the time of the patient's first biopsy. The period of time considered for

the analysis was one month. If the patients perform the same test more than once in that month the average of that test results is calculated.

3.3.2 Building Patient/Features Matrix

To extract patterns and classify patients we need to transform the dataset into a suitable format. According to the selected attribute the needed data available in I-lab and Biopsy table.

Figure 18 shows the **Patient/features matrix,** in this matrix each column represents a feature and each row represents the medical test results for each individual patient. Each patient has only one row that represents his/her characteristics.



	id	mounth	year	type	T_BIL	CHE	GOT	GPT	TP
▸	1	8	1983	B	0.7	8.47	80	121	7.7
	2	3	1992	C	0.6	280	25	46	6.6
	3	3	1987	B	0.8	4.45	25	31	7.05
	5	3	1992	C	1	354	40	76	7.1
	16	2	1992	C	0.4	376	115	163	7.7
	19	3	1992	C	0.9	284.5	21.5	29.5	6.7
	27	3	1983	B	1.25	4.435	317.5	413.5	8
	33	4	1988	B	0.95	316.5	70.5	168	7.15
	35	5	1992	C	0.8	286	70	128	7.5
	36	4	1992	C	1.1	361	56	44	7.1
	37	6	1998	C	0.86667	252.66667	89.33333	127	6.86667
	42	7	2000	B	0.7	205.5	21.5	24.5	6.2
*	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Figure 18: Patients/Features matrix

3.4 Data Normalization

Data normalization is a necessary step when dealing with parameters of different units and scales. It transforms data values for different attributes into a uniform set of units or into a uniform scale.

Data normalization is essential in data mining process because it improves the accuracy and the efficiency of the mining technique. These techniques provide better results if the data that will analyzed is normalized.

Data transformation such as normalization may improve the accuracy and efficiency of mining algorithms involving neural networks, nearest neighbor, and clustering classifiers.

Such methods provide better results if the data to be analyzed have been normalized.

In this study each attribute was normalized by scaling its values so the new values will be in the range of 0 to 10; $0 \leq \text{value} \leq 10$.

There are many methods for data normalization which includes min-max normalization, z-score normalization, and normalization by decimal scaling. The normalization technique that we apply in this work is min-max normalization

3.4.1 Min-Max Normalization

Min-max normalization performs a linear transformation on the original data.

Min-max normalization maps a value v of A to v' in the range **[new-mina, new-maxa]** by computing:

v'= ((v-mina) / (maxa – mina)) * (new-maxa – new-mina) + new-mina (18)

where_ **mina** and **maxa** are the minimum and the maximum values for attribute **A** and **v** is value for attribute **a**.

In this work, normalization process has done on Hepatitis dataset using min-max normalization. The values of the tests were normalized to be in [0, 10] range

3.5 Clustering Patients Using K-means

After building the patient/features matrix, the matrix was used to cluster users depending on tests results. Before clustering, patients' MID and hepatitis type were omitted from the matrix.

Multiple distance functions were applied on the prepared data sets which are square Euclidean, City block, Cosine and distance functions. Silhouette was used to indicate how the clusters are separated. The results are shown in chapter four.

3.6 Neural Network Architecture

ANN performs quit well with complex tasks such as classification problems using abstract data; therefore, this thesis applied ANN for classifying the hepatitis patients where in this thesis:

- A Feed forward Back propagation neural network is used.
- The number of neurons in the input layer is 5 (representing number of features for each sample).
- The number of neurons in the output layer is 1 (representing one class of the diseases the neural will generate).
- The training algorithm that was used for training the models is Resilient Back Propagation.
- The activation function that is used for all layers is TANH.
- The following values were used; 0.07 and 0.0000001 as the initial delta and performance goal error respectively.

3.6.1 Input layer

In this study, it is aimed to classify chronic hepatitis B and C patients using different tests. These tests were results from (Cheng et. al, 2006) who discovered that these tests clearly distinguish hepatitis B and C patients. In this study, five input variables essential to classifying hepatitis diseases of interest are extracted from hepatitis diseases database. The values of these tests are numerical variables, as mentioned in section 3.4. These values were normalized using min-max normalization to be in the range of 0 to 10.

3.6.2 Number of Hidden Layers and Hidden neurons

Several researchers have proposed some general rules for determining an optimal number of hidden neurons for any application. Moreover Kavzoglu as cited in (Yan et al., 2006) has shown that in most situations, there is no way to determine the best number of hidden layers without training several networks and estimating the generalization error of each, because the best number of hidden neurons depends mainly in a complex way on; the number of input and output units, the number of training cases and the complexity of the classification to be learned (Xu and Chen, 2008).

Based on this, experiments are the only way to determine the best number of hidden neurons. Therefore, in this study multiple architectures were tested starting with one hidden neuron and ending with 40 neurons. Figure 19shows the results of the accuracy rate with number of hidden nodes. This figure shows that when the number of hidden nodes is 15 the highest accuracy rate which is 85.29% is reached.

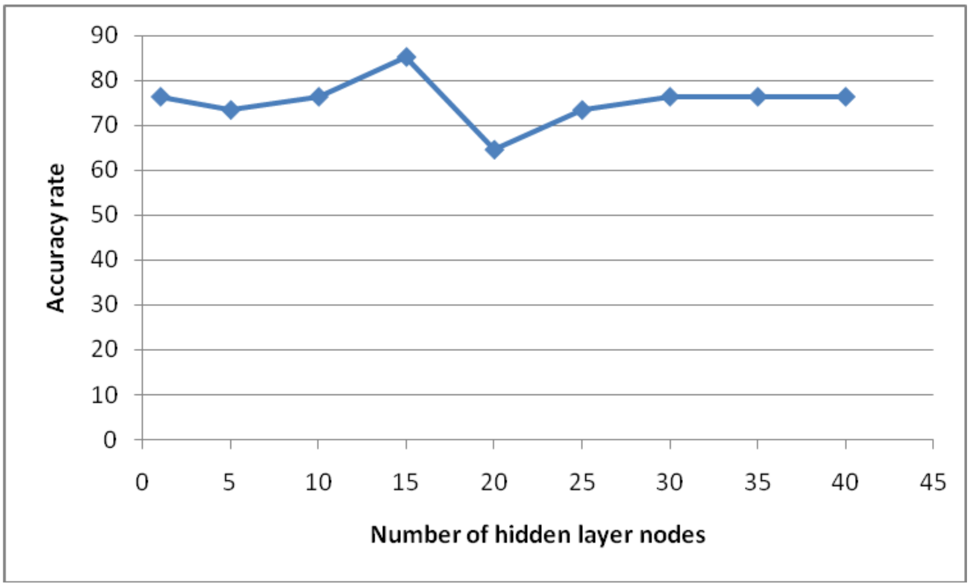


Figure 19: Accuracy rate vs. Number of hidden layer nodes

The final architecture that was used in this thesis is illustrated in Figure 20.

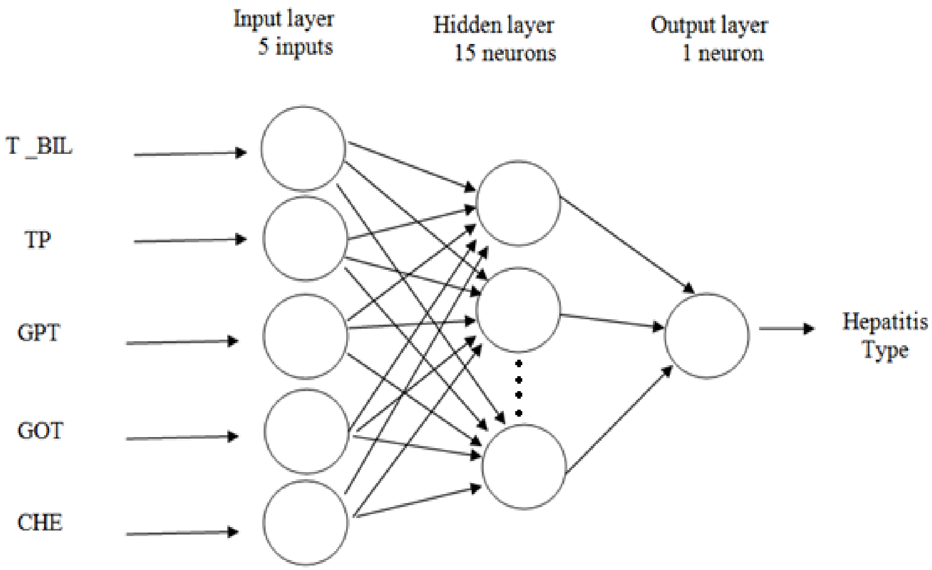


Figure 20: Neural network Architecture

4. Experimental Results and Discussion

4.1 Chapter Outline:

This chapter illustrates the results obtained from experiments which have been performed. It organized as follows: section one describes Hepatitis data set in details. Section two represents the experimental methodology; section three discusses the experimental results of the study for both K means and ANN. A comparison between the results obtained from K mean and ANN is represented in section five.

4.2 Hepatitis Data Set

This work used hepatitis data set that was offered as a part of discovery challenge at (Knowledge discovery in databases) PKDD 2005. Hepatitis data set, produced by Chiba University Hospital in Japan, contains results of laboratory tests for 771 patients with hepatitis B and C who underwent testing between 1982 and 2001.

The data set is in comma separated format (CSV) format. Comma Separated Value (CSV) file format is a common structuring strategy for text format files. In CSV files, each line in the file represents a row of data and commas in a line separate fields in the tables' row. Figure 4-1 depicts the main tables involved in the hepatitis data set.

The next subsections will explain each one of these tables.

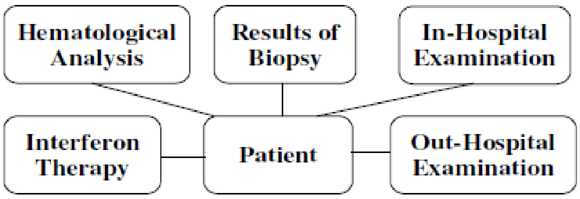


Figure 21: Hepatitis dataset tables (Pizzi et.al., 2005)

4.2.1 Patients Table

The Patients Table in the hepatitis data set contains the basic information of each patient such as the patient ID, sex and the date of birth, it should be noticed that the patient's ID is a key attribute that corresponds to the actual patient ID modified for privacy protection. Patients table contains 771 patients. Table 1 below shows each item and its meaning in the Patient Table.

Table1 : Patients Table

item	Meaning
MID	The ID of the patient
PT Sex	Sex of the patient
PT Birthdates	Birth date of the patient

4.2.2 Results of Biopsy Table

The Results of Biopsy Table in the hepatitis data set includes information about the biopsies results of each patient. It involves the MID of the patient who made the biopsy and hepatitis type, hepatitis subtype …etc. Table 2 below shows the items and its meaning in this table.

It is important to mention that not all patients had biopsies and the same patient may have more than one biopsy.

Table 2:Results of Biopsy

Item	Meaning
MID	Identification of the patient (masked)
BIOPSY Lo_No	Identification of the specimen
BIOPSY Hepatitis_Type	Type of hepatitis
BIOPSY Exam_Date	The date when biopsy was performed
BIOPSY Hepatitis_Subtype	Subtype of Hepatitis
BIOPSY Facility	Name of the facility where the specimen was collected
BIOPSY Fibrosis	Biopsy report about progress of fibrosis
BIOPSY Activity	Biopsy report about activity of virus

4.2.3 Interferon Therapy Table

The Interferon Therapy Table in the hepatitis data set gives information about the patients who were treated by interferon therapy. It should be noticed that not all patients were treated with interferon therapy. Table 3 shows each item in this table and its meaning.

Table 3: Information on interferon therapy

Item	Meaning
MID	Identification of the patient (masked)
IFN Administration_No	ID for each of the repeated administrations
IFN Administration_Date_ST	Start date of the administration
IFN Administration_Date_ED	End date of the administration

4.2.4 Out-Hospital Examinations Table

The Out-Hospital Examinations Table stores the results of tests that were done outside the hospital. Table 4 explains each item in the out-hospital table.

Table 4: Out-hospital examinations

Item	Meaning
MID	Identification of the patient (masked)
OLAB Exam_date	Date of the examination
OLAB Exam_No	ID for each examination performed repeatedly on the same day (n-th examination)
OLAB Exam_Name	Name/code of the examination
OLAB Exam_Condition	Comments on condition of the examinations
OLAB Exam_Result	Result of the examination
OLAB Unit	Unit for the measurement
OLAB Qualitative_Interp	Qualitative interpretation of the result
OLAB Comment1	Comments on the result
OLAB Comment2	Comments on the result
OLAB Evaluation	Evaluation of the result
OLAB Eval_SubCode	Internal subcode of the evaluation items

4.2.5 In-Hospital Examinations Table

The In-Hospital Examinations Table stores the results of exams that were done inside the hospital. Table 5 describes in-hospital examinations items.

Table 5: In-hospital examinations

Item	Meaning
MID	Identification of the patient (masked)
ILAB Exam_Date	Date of the examination
ILAB Exam_No	ID for each examination performed repeatedly on the same day (n-th examination)
ILAB Exam_Name	Name/code of the examination
ILAB Exam_Result	Result of the examination

4.2.6 Results on Hematological Analysis:

The Hematological Analysis Table gives the results of hematology analysis for the patients. Table 6 represents the items in the hematological analysis and the meaning of each one of them**.**

Table 6: hematological analysis

Item	Meaning
MID	Identification of the patient
HEMAT DATE	Date of the examination
HEMAT SEQ	ID for each of the repeated administrations
HEMAT WBC	White blood cells count
HEMAT RBC	Red blood cells count
HEMAT HGB	Hemoglobin count
HEMAT HCT	Hematocrit count
HEMAT MCV	Mean corpuscular volume
HEMAT MCH	Mean corpuscular hemoglobin
HEMAT MCHC	Mean corpuscular hemoglobin concentration
HEMAT PLT	Platelets count

4.3 Experimental Results

The different ANN architectures have been built and tested using MATLAB® version 2009a, which was run under windows Vista Home Premium SP2 OS with Intel (R) core (TM)2 Duo CPU of 2.10 GHz, and 3.00 GB of RAM. Moreover, the data set tables from the hepatitis data set were downloaded into SQL server database version 2005 and after applying the preprocessing phase (which was discussed in Chapter 3) a data set with 340 patients were produced with 170 patients for each type of the disease.

4.3.1 Classification Accuracy

For all experiments the evaluation metric that was used is the classification accuracy; The classification accuracy depends on the number of samples correctly classified and is evaluated by the formula:

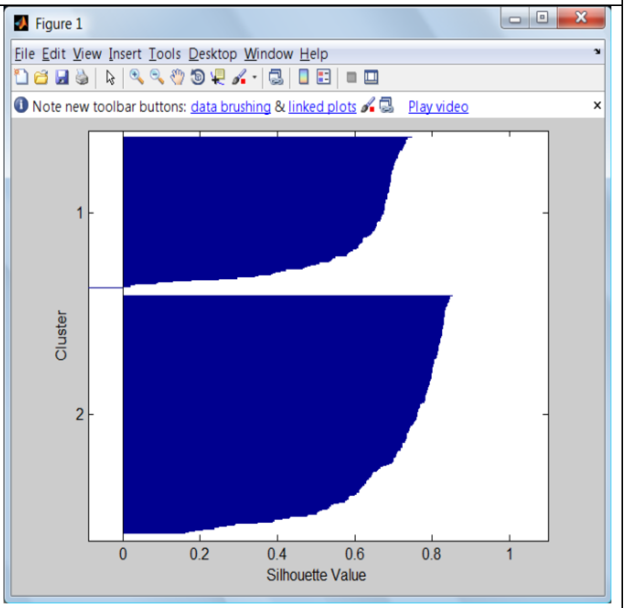
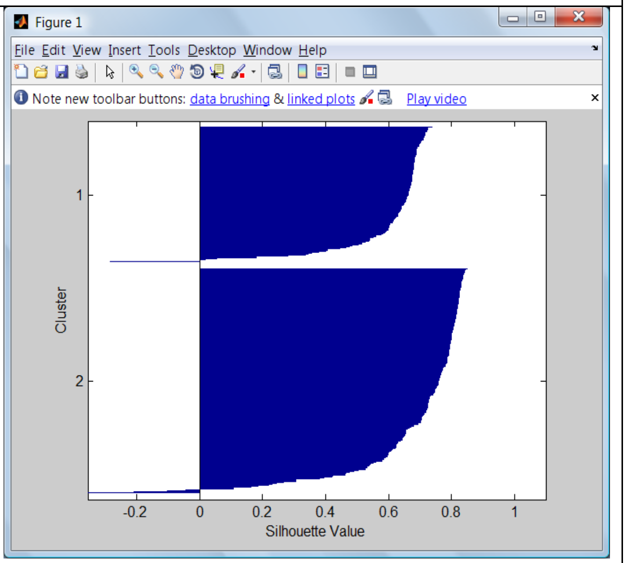
$$\textbf{Accuracy} = (t/n) * 100\%$$

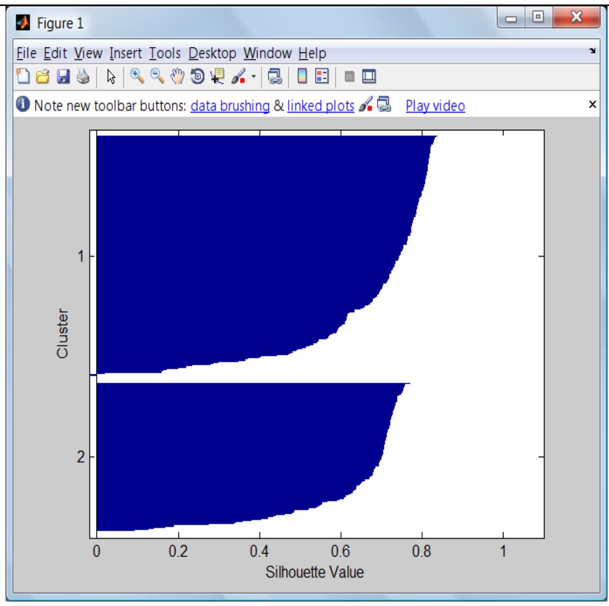
Where ***t*** is the number of sample cases correctly classified, and ***n*** is the total number of sample cases.

4.3.2 Patients Clustering Using K-Means

K means clustering was applied on the produced data set. Many distance functions were applied on the dataset. Silhouette was used to get an idea of how well-separated the clusters are. Table 7 shows silhouette plots generated for K-means clustering experiment with multiple distance functions.

Table 7: Silhouette Plots for Clustering with Various Distance Functions

<table>
<tr><th>Distance Function</th><th>Sillhoute Plot</th></tr>
<tr><td>Square Euclidean</td><td>
</td></tr>
<tr><td>Cosine</td><td>
</td></tr>
</table>

City block	

It was noticed that the silhouette plots in table 7 almost give the same results, both of them shows that the patients are well distributed.

4.3.2.1 Validation Patients Clustering Using K-Means

It is necessary to analyze the above silhouette in order to justify the results obtained. This is done through the following steps

1. The correct hepatitis type for each patient was obtained from patients/features matrix.
2. Comparison between k means result and patients/features results for each patient were made.

Finally, accuracy for the k mean algorithm was calculated. The classification accuracy depends on the number of samples correctly classified.

Table 8 shows the results analysis for the above silhouettes

Table8 : Clustering analysis

Distance function	Number of patients		Number of correctly classified patients	
Squared Euclidean distance function	Cluster 1 209	Cluster 2 131	Cluster 1 162	Cluster 2 124
Cosine distance function	Cluster 1 128	Cluster 2 212	Cluster 1 123	Cluster 2 165
City Block distance function	Cluster 1 130	Cluster 2 210	Cluster 1 123	Cluster 2 163

Table 9 represents the classification accuracy obtained Squared Euclidean, Cosine and city block distance functions.

Table9 : Classification Accuracy

Distance Function	Total Number of Patients	Total Number of Correctly Classified Patients	Accuracy
Squared Euclidean Distance Function	340	286	(286/340)*100% =84.11%
Cosine Distance Function	340	288	(288/340)*100%= 84.70%
City Block distance function	340	286	(286/340)*100%= 84.11%

The highest accuracy obtained was 84.70 using Cosine Distance function. This result considered to be good for the following reasons:

- The nature of the data; the data set is medical data set and input mistakes is possible.
- Many preprocessing stage were done and that is reflect the values of the attributes.

4.3.3 Results of Applying ANN Technique

One of the major advantages of neural networks is their ability to generalize; this means that a trained network could classify data that it has never seen before from the same class as the learning data. In real world applications developers normally have only a small part of all possible patterns for the generation of a neural network, therefore, to reach the best generalization, the dataset should be split into two parts:

- **The training set** that is used to train a neural network. The error of this dataset is minimized during training.
- **A test set** for finally checking the overall performance of a neural network.

In this work, 90% of the data was used as the training data and the remaining 10% of the data was used for the testing process. Table 10 shows how the data set was divided into training and testing sets.

Table 10: ANN Data Set

Data set	**Total number of patients**	**percentage**	**Number of patients with Hepatitis type B**	**Number of patients with Hepatitis type C**
Training set	300	90%	150	150
Testing set	34	10%	17	17

In ANN the training should be stopped in the minimum of the test set error. At this point the network generalizes best. When learning is not stopped overfitting occurs and the performance of the network on the whole data decreases. So, in this work the network was tested with the test set for many epochs (times) in order to investigate when the training should be stopped. The best classification accuracy obtained was 85.29% when the network was trained for 350 epochs . The classification accuracies that were obtained with different epochs are illustrated in Table 11 and Figure 22. As noticed from the results, the accuracy started to increase until reaching 350 epochs after that an overfitting occurred where the accuracy started to decrease and continued decreasing.

Table 11: ANN results

Number of epochs	Accuracy
100	76.47
150	76.47
200	82.35
250	82.35
300	85.29
350	85.29
400	82.35
450	79.41
500	79.41
550	79.41
600	79.41

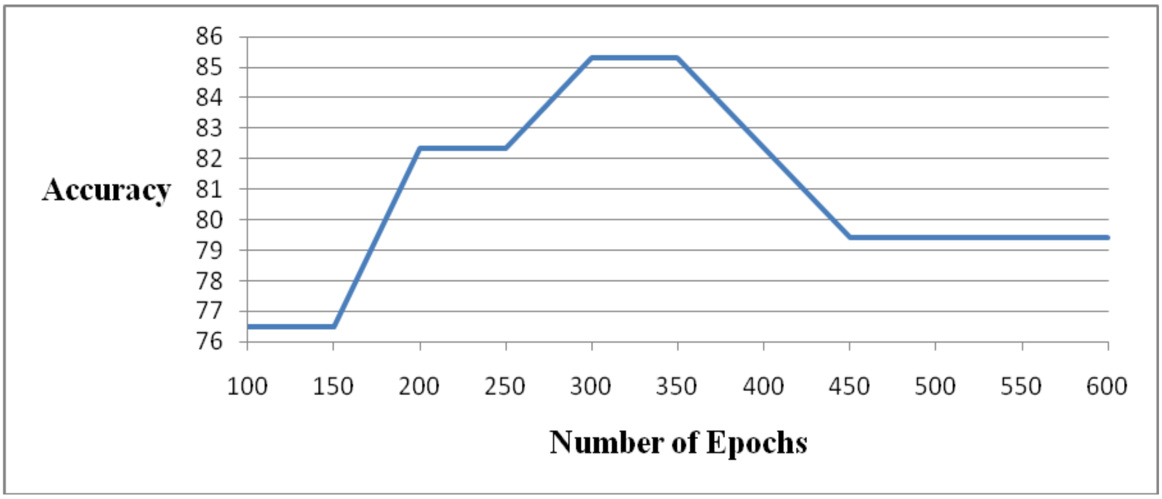


Figure 22: ANN Classification Accuracies Curve.

The experimental result of the best classification accuracy that was obtained has been presented as a confusion matrix in Table 12; a confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such a system is commonly evaluated using the data in the matrix. In a confusion matrix, columns represent the predicted data, while rows represent the actual data.

Table 12: Confusion Matrix

Disease Type	Hepatitis Type B	Hepatitis Type C	Row Sum
Hepatitis Type B	13	4	17
Hepatitis Type C	1	16	17
Column Sum	**14**	**20**	**34**

4.3.3.1 Justification ANN Technique Results

Form the confusion matrix shown in Table 12 the accuracy for each class can be calculated as presented in Table 13; where the proportion of the disease type in the test dataset that have been correctly recognized by the ANN is shown.

Table 13: Proportion of Correctly Classified Diseases

Disease Type	Accuracy %
Hepatitis B	**(13/17)*100%=0.76 %**
Hepatitis C	**(16/17)*100%=0.94 %**
Average accuracy	**(1.71/34)*100%=85.29 %**

We notice that the hepatitis B disease obtained the lowest classification accuracy with a percentage of 76%, the highest classification accuracy was obtained by the hepatitis C with a percentage of 94%, giving an average accuracy of 85.29%.

We due these results to the following reasons:

- The data set is not that big and because of that not all cases may available
- Hepatitis C considered more stable in status than B.

4.3.4 Summary of Results

In this study two classification techniques were used which are k means and ANN.

Table 14 shows summary of results obtained by both K-means and ANN.

Table 14: Summary of Results

Classification Technique	Classification Accuracy obtained
ANN	85.29 %
K-means	84.70

5. Conclusion and Future Works

5.1 Conclusion

In this study chronic hepatitis patients with B and C was classified into clusters based on results of their tests. The main problem faced is choosing the right interval and selecting important features. Normalization and cleansing techniques were performed.

After data set preparation two classification techniques were applied; the first technique is K means which is supervised learning. The accuracy that was obtained is 84.70%.

The other classification technique is ANN which is supervised learning technique. The details on how we have realized, designed, and tuned the system was provided which was built based on a MLP feed forward neural network and trained using the Rprop training algorithm. It consisted of 5 input variables, one hidden layers, where the number of neurons in the hidden layer was 15. The output layer produced the hepatitis type for the patient. We train the network number of time until we get the highest accuracy. Which is 85.29 %.

5.2 Future Work

In future, this work may be extended using other types of ANNs such as GRNN, PNN and RBF and compare the results obtained from them with the one obtained using the MLP. Also different distance functions may be applied. Furthermore; another medical tests may be chosen.

REFERENCES

Bascil, S. And Temurtas, F. (2009), A Study on Hepatitis Disease Diagnosis Using Multilayer Neural Network with Levenberg Marquardt Training Algorithm, **Journal of Medical Systems,** 33 (1), 1-4

Caio, S., Philicity, W., Juan, E. And Gerry, D. (2010), A class-specific ensemble feature selection approach for classification problems, **In Proceedings of the 48th Annual Southeast Regional Conference (ACM SE '10)**, New York, USA, 1-6

Cheng, T., Wei, C., Tseng, V. (2006), Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches, **In Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06),** Utah, USA, 165-170

Dash, M., and LIU, H. (1997), Feature selection for classification, **Intelligent Data Analysis,** 1,131–156

Dienstag, JL., Isselbacher, KJ., Braunwald, E. and Kasper, DL. (2001), Acute viral hepatitis, **Harrison's Principles of Internal Medicine.** New York: McGraw-Hill

Duda, R., Hart, P. and Stork, D. (2001), **Pattern Classification**, (2nd ed), NewYork: John Wiley & Sons, Inc.

Fausett, L., Laurene, M. (2001), **Fundamentals of Neural Networks**, **Architectures, Algorithms and Application**,(1st ed), New Jersey: Pearson Prentice Hall Education , Inc.

Fayyad, U., Shapiro, G., and Symth, p. (1996), From Data Mining to Knowledge Discovery in Databases. **Artificial intelligence Magazine**. 17, 37-54

Freeman, J., and Skapura, D. (1991), **Neural Networks Algorithms, Applications, and Programming Techniques,** (1st ed), New York: Addison-Wesley Publishing Company, Inc

Geoffrey, H. and Terrence, S. (1999) **Unsupervised Learning: Foundations of Neural Computation** (1st ed), Massachusetts: MIT Press**.**

Hagan, M., Dcmuth, H. and Beale, M. (1996), **Neural Network Design,** (1st ed), USA: PWS Publishing Company.

Han, J. and Kamber, M., (2000) **Data Mining: Concepts and Techniques,** (1st ed), Morgan Kaufmann.

Haykin, S. (2001), **Neural Networks a comprehensive foundation**, (2nd ed), India: Pearson Prentice Hall Education, Inc.

Hirano, S. and Tsumoto, S. (2005), Clustering Time-Series Medical Databases Based in the Improved Multiscale Matching In: Hacid, Mohand-Said , Murray, Neil V., Ras, Zbigniew W. and Tsumoto Shusaku (Ed), **Foundations of Intelligent Systems** (pp 612-621), Berlin: Springer,

Hirano, S. and Tsumoto, S. (2006), Cluster Analysis of Time-series Medical Data Based on the Trajectory Representation and Multiscale Comparison Techniques, **In Proceedings of the Sixth International Conference on Data Mining (ICDM'06),** Hong Kong, China, 896 - 901

Hirano, S. and Tsumoto, S. (2007), Temporal Data Mining in Hospital Information systems: Analysis of clinical courses of chronic Hepatitis, **International Journal of Intelligent Computing,** 1 (1), 11-19

Hirano, S. and Tsumoto, S. (2008), Granularity in Temporal Data Mining, **In Proceeding of Information Processing in Management of Uncertainty in Knowledge-Based Systems** MU'08, Malaga, Spain, 1179-1785.

Ho, T., Yokoi, H., Takabayashi, K., Kawasaki, s., Le ,Nguyen, D. (2003), Mining Hepatitis Data with Temporal Abstraction. **In Proceeding of the ACM International Conference on Knowledge Discovery and Data Mining KDD’03**, New York, USA, 369–377.

Jain, A. and Dubes, R. (1988), Algorithms for Clustering Data. **Prentice Hall, Englewood Cliffs, NJ.**

Jain, Anil., and Douglas Z., (1997), Feature selection: Evaluation, application, and small sample performance. **Pattern Analysis and Machine Intelligence,** 19(2), 153–158.

Kenji, S., Okubo, Y., Haraguchi, M., and Kunifuji1, S. (2007), "Data Mining of Time-Series Medical Data by Formal Concept Analysis", **Springer-Verlag Berl in/ Heidelberg** , Part II, LNAI 4693, pp. 1214–1221.

Krose, B., and Smagt, P. (1996), **An introduction To Neural Networks**, (8th ed), Amsterdam: The University of Amsterdam.

Larose, DT. (2006), **Data Mining: Methods and Models**. New York: John Wiley & Sons,Inc.

Lei, Y. and Huan Liu. (2004), Efficient Feature Selection via Analysis of Relevance and Redundancy, **Journal of Machine Learning Research**. 5, 1205-1224.

Lemke, F. and Mueller, J.(2003), Medical data analysis using self-organizing data mining technologies. **Systems Analysis Modeling Simulation**. 43(10), 1399-1408.

Leung, K., Lee, K., Wang, J., Ng, T. , Tsui, S. , and Mok, S. (2011), Data Mining on DNA Sequences of Hepatitis B Virus , **Computational Biology and Bioinformatics,** 8, 428-440

MacQueen, B. (1967), Some Methods for classification and Analysis of Multivariate Observations. **In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.** California, USA, 281–297

Martine Swell (2007), **Machine Learning**, A technical report, Department of computer science, University College London, London, UK.

Mehrotra, K., Chilukuri, K., Mohan and Ranka, S. (1996) ,**Elements of Artificial Neural Networks**,(1st ed), New York: Addison-Wesley Publishing Company, Inc.

Mitchell, T., (1997), **MachineLearning**, (1st ed), McGraw-Hill Science.

Ozyilmaz, L., and Yildirim (2002), Diagnosis of the thyroid disease using artifiacial neural networks methods. **In Proceeding of the 9th International Conference on Neural Information Processing**, Istanbul, Turkey, 2033 - 2036

Peters, G., (2007), Diagnosis and Management ofHepatitis B Virus and HIV Coinfection**. IAS-USA. Topic in HIV Med.,**15:163-166

Pizzi, L., Ribeiro, X. and Vieira M. (2005), Analysis of hepatitis dataset using multi-relational association rules. **In Proceeding of the ECML/PKDD, Discovery Challenge,** Berlin, Germany, 161-167.

Ramalingam, V., Palaniappan, B., Panchanatham, N. and Palanivel, S. (2006), Measuring Advertisement Effectiveness – A Neural Network Approach, **Expert Systems with Applications,** 31(1), 159–163.

Riedmiller, M. and Braun, H. (1993), A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm, Neural Networks**, IEEE International Conference, San Francisco, CA , USA** ,586-591.

Thammano, A., Moolwong, J. (2009), A New Computational Intelligence Technique Based on Human Group Formation, **Expert Systems with Applications**, 37(2), 1628–1634

Thapa, B.R. and Walia , A. (2007), Liver Function Tests and their Interpretation, **Indian Journal of Pediatrics**, 74, 663-671.

Xu, S. and Chen, L. (2008), A Novel Approach for Determining the Optimal Number of Hidden Layer Neurons for FNN's and Its Application in Data Mining**, In the 5th International Conference on Information Technology and Applications,** Cairns, Australia, 683-686

Yan, H., Jiang, Y., Zheng, J., Peng, C. and Li, Q. (2006), A Multilayer Perceptron-Based Medical Decision Support System for Heart Disease Diagnosis, **Expert Systems with Applications**, 30(2), 272-281.

Yoon, S. C., Henschen, L. J., Park, E. K., and Makki,S. (1995), Using Domain Knowledge in Knowledge Discovery, **Proceedings of Eighth International Conference on Information and KnowledgeManagement**, pp.243-250.

Zhang, G. and Berardi V. L.,(1998), An investigation of neural network in thyroid function diagnosis , **Health Care Management Science,** 1, 29-37

URL: http://lisp.vse.cz/challenge/ Retrieved May 15, 2010

Appendix A: Patients Table

MID	PT Sex	PT BirthDate
1	M	19540304
2	M	19590220
3	M	19481109
4	M	19380301
5	M	19510418
8	M	19551214
12	M	19440916
16	F	19340604
19	M	19300623
20	M	19291217
21	M	19270301
27	M	19570101
29	M	19350714
30	F	19251001
32	F	19480124
33	M	19550713
35	M	19550804
36	M	19240808
37	M	19461215
39	M	19381116
40	F	19621110
41	F	19620310
42	F	19740725
45	F	19540220
46	F	19370508
47	F	19111125
49	F	19331101
50	F	19441201
51	F	19611109
52	F	19660511
53	F	19730316
54	F	19691129
55	F	19680823
56	F	19650406
57	F	19760113
58	F	19670705
9	F	19740625

Appendix B: Biopsy Table

MID	BIOPSY Lo_No	BIOPSY Hepatitis_Type	BIOPSY Exam_Date	BIOPSY Hepatitis_Subtype	BOIPSY Facility	Biopsy Fibrosis	Biopsy Activity
1	2181	B	19830819			2	2
2	5484	C	19920319			1	1
3	3781	B	19870305			1	1
5	5479	C	19920319			1	1
16	5437	C	19920220			4	2
19	5491	C	19920326			1	2
19	8007	C	19960307			1	0
27	1949	B	19830303			2	3
33	4177	B	19880428			2	1
35	5574	C	19920518			1	1
36	5534	C	19920425			1	1
37	8900	C	19980611			3	2
42	9393	B	20000704			1	2
45	8700	B	19970826			4	3
49	5625	C	19920609			1	1
50	5692	C	19920702			1	2
51	9413	B	20000809			3	2
54	2264	B	19831102			1	1
56	2248	B	19831027			1	1
57	7217	B	19940823			1	2
58	4195	B	19880520			2	2
59	7376	C	19950112			1	1
60	4098	B	19880129			2	1
61	7695	B	19950815			3	3
62	6175	C	19930213			1	
66	3467	B	19851212			1	1
69	2224	B	19830928			1	1
70	7966	B	19960119			1	2
71	5052	C	19901101			0	1
71	5098	C	19901227			0	0
72	4782	C	19900111			1	1
72	5272	C	19910718			1	1
72	5882	C	19920918			1	1

Appendix C: Interferon Therapy Table

MID	IFN Administration_No	IFN Administration_Date_ST	IFN Administration_Date_ED
2	1	19920615	19921129
16	1	19920305	19920527
19	1	19920608	19921120
35	1	19920706	19921218
36	1	19921028	19930215
37	1	19980807	19990123
49	1	19921027	19921208
50	1	19920811	19921128
59	1	19950222	19950419
71	1	19901105	19901216
72	1	19920919	19921204
78	1	19901203	19910113
86	1	19951011	19960228
87	1	19990714	19991209
91	1	19920910	19921111
104	1	19950323	19950612
111	1	19960416	19961011
122	1	19930927	19931210
123	1	19880815	19890815
125	1	19940928	19950327
133	1	20000928	20010326
138	1	19931018	19940415
142	1	19961104	19970501
148	1	19921124	19930510
149	1	19911015	19920420
158	1	19930506	19930906
163	1	19871116	19880108
169	1	19940831	19950228
171	1	19951016	19951120
176	1	19901126	19910106
179	1	20010322	20010905
180	1	19950412	19950523
181	1	19920701	19921216
182	1	19920924	19930327
183	1	19921110	19930427

Appendix D: Out-Hospital Examinations Table

MID	OLAB Exam_Date	OLAB Exam_No	OLAB Exam_Name	OLAB Exam_Condition	OLAB Exam_Value	OLAB Unit	OLAB Qualitative_Interp	OLAB Comment2	Field 10	OLAB Evaluation	OLAB Eval_SubCode
1	19850711	1	CA19-9		8	U/ML					
1	19870114	1	CMV.IGG(ELISA)		0.729		(2+)				
1	19870114	1	CMV.IGM(ELISA)		0.214		(-)	saikenzu midesu			
1	19870513	1	DNA-porimera-ze		14	CPM					
1	19870903	1	AFP		1.22	NG/ML					
1	19870903	1	HBE-AB		100		+				
1	19870903	1	HBE-AG		0.3		-				
1	19870903	1	HBS-AG		36.9		+				
1	19871202	1	AFP		1.75	NG/ML					
1	19871202	1	HBE-AB		100		+				
1	19871202	1	HBE-AG		0.3		-				
1	19880203	1	AFP		1.31	NG/ML					
1	19880203	1	HBE-AB		100		+				
1	19880203	1	HBE-AG		0.3		-				
1	19880511	1	AFP		1.4	NG/ML					
1	19880511	1	HBE-AB		100		+				
1	19880511	1	HBE-AG		0.5		-				
1	19880608	1	HBE-AB		100		+				

Appendix E: In-Hospital Examinations

MID	ILAB Exam_Date	ILAB Exam_No	ILAB Exam_Name	ILAB Exam_Result
213	19961106	1	A2PI	93
230	19940726	1	A2PI	44
230	19940628	1	A2PI	56
230	19940705	1	A2PI	40
230	19940704	1	A2PI	48
205	19910516	1	A2PI	22
911	19920930	1	A2PI	77
911	19930128	1	A2PI	54
911	19930118	1	A2PI	51
911	19930215	1	A2PI	48
911	19930305	1	A2PI	59
911	19930222	1	A2PI	54
888	19960704	2	A2PI	67
888	19960627	2	A2PI	48.0
888	19960708	1	A2PI	74.0
548	19941226	1	A2PI	103
548	19930921	2	A2PI	93
876	19940705	1	A2PI	82
672	19930602	1	A2PI	91
253	19961024	1	A2PI	63
253	19961028	1	A2PI	59
253	19961015	1	A2PI	47.0
807	19990129	2	A2PI	81
120	19930531	1	A2PI	61
120	19930524	1	A2PI	56
120	19930513	1	A2PI	59
744	19980615	2	A2PI	63
943	19960130	1	A2PI	27
943	19960201	1	A2PI	25
632	19931008	1	A2PI	99
239	19930729	1	A2PI	72
609	19990712	2	A2PI	88
903	19950904	1	A2PI	46
585	19980811	2	A2PI	85
904	19920612	1	A2PI	70
928	19980116	2	A2PI	114

Appendix F: Hematological Analysis Table

MID	HEMAT DATE	HEMAT SEQ	HEMAT WBC	HEMAT RBC	HEMAT HGB	HEMAT HCT	HEMAT MCV	HEMAT MCH	HEMAT MCHC	HEMAT PLT
230	19841022	1	2.5	4.66	14.3	42.8	91	30.6	33.4	77
230	19841029	1	2.2	4.27	13.4	38.6	90	31.5	34.8	60
230	19841101	1	3.4	4.18	12.7	38.6	92	30.4	33	
230	19841105	1	7.9	4.38	13.6	40.1	91	31	33.9	53
230	19841114	1	8	3.96	12.2	35.8	90	30.8	34	162
230	19841119	1	2.4	3.33	10.2	30.3	90	30.9	33.9	139
230	19841122	1	2.9	3.67	11.3	33.4	91	30.8	33.9	131
230	19841124	1	1.8	3.44	10.7	31.6	91	31.1	33.9	103
230	19841203	1	2	3.8	11.7	35.2	92	30.9	33.4	67
230	19841208	1	2.2	3.47	10.9	31.9	92	31.5	34.2	48
230	19841213	1	1.8	3.56	10.9	33.1	92	30.7	33	52
230	19841222	1	1.4	3.66	11.1	33.1	90	30.5	33.6	65
230	19850105	1	1.3	3.81	11.5	34.9	91	30.3	33.1	66
230	19850111	1	2	4.23	12.8	38.9	92	30.2	32.8	84
230	19850118	1	1.5	3.9	12.1	35.6	91	31	34	79
230	19850124	1	1.5	4.04	12.3	37.2	92.1	30.4	33.1	71
230	19850130	1	1.9	4.12	12.9	37.9	92	31.3	34	57
230	19850228	1	2.1	4.02	12.4	38	94.5	30.8	32.6	41
230	19850425	1	2.2	4.04	12.6	38.3	94.8	31.2	32.9	49
230	19850912	1	2.4	4.04	13	39.2	97	32.2	33.2	45
230	19851018	1	1.9	3.99	12.9	38.5	96.5	32.3	33.5	44
230	19851022	1	1.8	3.48	11.2	33.4	96	32.2	33.5	29
230	19851024	2	1.4	3.33	11	32.2	96.7	33	34.2	34
230	19851025	1	1.8	3.5	11.4	33.6	96	32.6	33.9	38
230	19860109	1	2.4	3.96	12.6	38.1	96.2	31.8	33.1	63
230	19860515	1	2.3	4.03	12.7	37.9	94	31.5	33.5	45
230	19861009	1	2.7	3.82	12.4	36	94.2	32.5	34.4	43
230	19870702	1	3	4.37	13.8	41.2	94.3	31.6	33.5	38
230	19870820	1	2.6	4.34	13.8	41.5	95.6	31.8	33.3	42
230	19880128	1	2.6	4.19	13.4	40	95.5	32	33.5	51
230	19880414	1	2.6	4.05	13.1	39	96.3	32.3	33.6	54
230	19880616	1	3.1	4.04	13.2	39	96.5	32.7	33.8	51
230	19881124	1	3.1	4.02	13.2	38.2	95	32.8	34.6	61

تشخيص مرض الكبد الوبائي باستخدام تقنيات استخراج البيانات

إعداد
منال عبد الله عبد عبد الجبار

المشرف
الدكتور عمار الحنيطي

المشرف المشارك
الأستاذ الدكتور عزمي محافظة

ملخص

التهاب **الكبد الوبائي** هو احد انواع الالتهابات التي قد تصيب الكبد حيث يسبب هذا النوع من الالتهاب فيروس معين يرمز له باحد الاحرف A – B – C – D – E.
أشارت الإحصائيات الطبية أن حوالي 3% من سكان العالم أي أكثر من 170 مليون شخصاً يعانوا من الإصابة بمرض صامت يقوم بتدمير **الكبد.**

تقدم هذه الدراسة إطار متكامل يهدف إلى تصنيف مرضى الكبد الوبائي الفيروسي الى نوعية ب و ج . لقد تم استخدام قاعدة بيانات مقدمة من مستشفى شيبا في اليابان والتي تحتوي على النتائج المخبرية ل 771 مريض مصابين بمرض الكبد الوبائي المزمن ب و ج.

تحتوي قاعدة البيانات المستخدمة على كمية هائلة من البيانات مما تتطلب القيام بالعديد من العمليات للحصول على البيانات المطلوبة بالصيغة المناسبة . ناتج تلك العمليات كان 340 مريض 170 مريض لكل نوع من الكبد الوبائي.

بعد ذلك تم تصميم نظاما لتحديد نوع الكبد الوبائي بواسطة شبكة عصبية اصطناعية متعددة الطبقات ذات تغذية أمامية. تستقبل طبقة الإدخال 15 متغير. في حين أن طبقة الإخراج ستنتج نوع المرض .لقد تم استعمال طريقة تكرارية لتحديد عدد الطبقات المخفية و عدد الخلايا العصبية في هذه الطبقات. لتدريب الشبكة تم استعمال خوارزمية التدريب (RProp).
وقد أظهرت التجارب بأن النظام المقترح في هذه الرسالة استطاع التنبؤ بنوع مرض الكبد الوبائي بدقة 85.29% .